Development of spatio-temporal data fusion frameworks for point and gridded soil moisture data

Weiyue Zheng

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Mathematics and Statistics
College of Science and Engineering
University of Glasgow

December 2025

Abstract

Monitoring soil moisture can play an important role in helping to inform researchers, regulators, and landowners about the available water content of the soil for agriculture and vegetation. However, the capacity to observe soil moisture is constrained by practical and financial limitations, making it challenging to observe continuously across space and time. We can only monitor soil moisture at a finite number of spatial locations and time points. One of the most accurate methods for measuring soil moisture is to use in-situ sensors. However, the high cost of deploying these sensors extensively means that soil-moisture data tends to be collected from a sparse network of monitoring points.

Given the limited in-situ sensor data, it becomes essential to explore the benefits of utilising other data sources, such as satellite data, by developing and using data fusion techniques. Data fusion allows for the integration of different data sources, enhancing the ability to make informed decisions and understand environmental phenomena with more precision, despite the limited direct monitoring of soil moisture.

The research question is motivated by the in-situ soil-moisture data provided by SEPA in Elliot Water and the satellite images provided by Copernicus. It is necessary to develop a data-fusion method for point data and gridded data, so that the accuracy of the in-situ data can be combined with spatial and temporal information from satellite data to generate a fine-resolution map with uncertainty quantification.

This thesis introduces three INLA-based data-fusion methods in Chapter 3, 4, and 5, which include a spatio-temporal regression with misaligned covariates, a spatial data fusion method, and a spatio-temporal data fusion method. A comprehensive simulation study varies sensor density, grid resolution, percentages of missing grid data, and temporal window length k. Across scenarios, joint fusion consistently outperforms point-only and grid-only baselines in RMSE. This thesis also introduces an XGBoost-based constrained ensemble method with conformal prediction in Chapter 6, developed to merge in-situ point and satellite gridded data under different spatio-temporal supports.

This thesis presents the background, motivation, model development, and application of the novel data fusion methods, addressing the gap in the literature by accounting for spatio-temporal change-of-support problems. Results are presented throughout to demonstrate the use of the data fusion model in soil moisture data.

Contents

List of Tables

List of Figures

Acknowledgements

I want to thank everyone who supported me during my PhD journey.

First, thank you to my supervisors, Professor Marian Scott, Professor Claire Miller, and Dr Andrew Elliott, for always encouraging me, guiding me, and sharing your knowledge. I'm very grateful for your patient reading of my rough drafts, for listening to my half-formed ideas, and for giving helpful feedback that lifted my work to new heights.

Thanks to the School of Mathematics and Statistics at the University of Glasgow for giving me a great place to work and the computing resources I needed. I especially appreciate the IT team for their quick help with the high-performance computing cluster and other technical issues.

I'm lucky to have so many kind colleagues and friends. Thank you to former PhD students Robin Muegge, Danniela Cuba, and Wenhui Zhang, and to Craig Wilkie, for early conversations that pointed me in the right direction, especially at RSS 2022 in Aberdeen, my first conference, which I'll always remember. Thank you also to Lanxin Li, Yuan Liu, Xueqing Yin, and everyone who stayed by my side during the pandemic. To my officemates, Steven Jun Villejo, Yizhu Wang, Yue Zhang, Mengran Li, and others, your morning "How are you?" brightened my day, and your help with INLA code was priceless.

To my friend Mahnoz, thank you for the fun breaks that kept me going, and to Dan Liu, your tips on thesis writing and your example of hard work gave me confidence. To my badminton and bouldering partners, your company reminded me to take breaks and stay balanced.

Above all, I owe everything to my parents. Your love, belief in me, and support have been my biggest strength.

Finally, thanks to everyone who mentioned here or not, who helped me finish this thesis. I couldn't have done it without you.

Declaration

I declare that I have finished this thesis by myself and that all work described in the thesis was carried out by me, except as otherwise clearly stated and referenced in the text. I confirm that this work has not been previously submitted for any degree or professional qualification. Part of Chapter 3 was published in the Proceedings of the 38th International Workshop on Statistical Modelling (IWSM), with the title "Spatial regression with misaligned covariates for soil moisture mapping", and was presented as an oral presentation at this conference. The content in Chapter 4 and 5 was presented as a talk at the CMStatistics conference in 2023. The content in Chapter 5 and 6 was presented as a poster at the Spatial Statistics conference in 2025. The manuscripts based upon the material in Chapter 4, 5, and 6 are currently in preparation.

I used ChatGPT only for language and minor coding suggestions. All research design, analyses, results, and interpretations are my own. I am entirely responsible for all content in this thesis.

Chapter 1

Introduction

This chapter introduces the background and motivation, summarises the study catchment and datasets, outlines the main statistical developments in data fusion, and gives a high-level overview of the concepts and methods used in this thesis. It ends with the research questions, contributions, and the thesis structure.

1.1 Background and motivation

Soil moisture is an important variable in hydrology, agriculture, and climate science. It affects how water moves through the land, controls how much water plants use, and plays an important role in weather and climate systems (Entekhabi et al., 1996; Seneviratne et al., 2010). Thus, monitoring soil moisture is critical for managing drought, predicting floods, planning irrigation, and improving climate models. However, getting an accurate and consistent soil moisture map across large scale areas is difficult. Ground-based sensors give precise measurements at specific locations, but they are sparse and do not cover wide spatial regions (Dorigo et al., 2011). On the other hand, satellite missions like Sentinel-1 provide broad spatial coverage, but the data have lower resolution (European Space Agency, 2025a), and they may be affected by clouds or vegetation (Ochsner et al., 2013a; Kerr et al., 2010). For Sentinel-1 specifically, it gives detailed and consistent spatial coverage for mapping soil moisture changes, but the soil moisture estimates are indirect and depend on incidence angle, surface roughness, vegetation, and topography, so products typically require filtering and aggregation (0.1 to 1 km). These two types of point-based and satellite-based data have different strengths and weaknesses. As a result, many recent studies have focused on combining them using data fusion methods to generate high-resolution soil moisture maps with uncertainty quantification. This research will take advantage of the detailed local accuracy of sensors and the wide spatial coverage of satellites. By integrating multiple sources, fusion methods can produce soil moisture maps that are both high-resolution and more complete over space and time (Gruber et al., 2019).

This thesis has several novelties. First, we directly address the challenge of spatially misaligned

covariates, which is common in spatial regression but often has been overlooked. Second, we develop a spatio-temporal INLA-SPDE data fusion framework that combines point sensors with satellite gridded data while properly accounting for misaligned covariates. Third, we introduce a constrained ensemble learning and conformal data fusion approach that provides uncertainty quantification, which explores classical statistical data fusion methods with modern machine learning models. On the application side, we present, to our knowledge, the first soil moisture data fusion for the Elliot Water catchment, combining in-situ point sensors with satellite gridded data to deal with the change-of-support problem. Together, these contributions advance both the methodological and application sides of data fusion and soil moisture mapping.

1.2 Study region and data sources

The study focuses on the soil moisture in the catchment area of Elliot Water (Angus, Scotland) in Figure 1.1, which is located in Angus, eastern Scotland. The site is of particular interest because of its agricultural land use, which contributes to the water quality management in this area. The catchment also has established in-situ sensor monitoring and broad satellite coverage, which provides multi-scale observations. These factors make Elliott Water a valuable case study for soil moisture monitoring and data fusion methods.

The study uses multiple data sources to uncover spatial-temporal patterns within the data, which include data from COSMOS soil moisture sensors (UK Centre for Ecology and Hydrology, 2024), Scottish Environment Protection Agency (SEPA) environmental monitoring sensors (Scottish Environment Protection Agency, 2024), and Copernicus satellite images (Copernicus Land Monitoring Service, 2024). Elevation and soil types (Open-Elevation, 2023; James Hutton Institute, 2024), which are two important variables impacting soil moisture, are considered in this study to support further modelling. This section will briefly describe the sensor data and satellite data used in this study, introduce the study catchment and the data scope, explain the details of each dataset, and describe how the data are collected. The details of the data sources are described below.

1.2.1 SEPA data

The SEPA maintains and operates the Elliot Water Live Sensors, which collect environmental and soil-related data. These sensors measure soil properties such as volumetric water content (VWC) (%), soil temperature (°C), and conductivity (dS/m), providing insights into soil moisture and conditions. VWC is a key variable in hydrology and soil science, representing the volume of water within a given volume of soil (Al Majou et al., 2008). It is usually expressed as a percentage or a decimal and is a critical indicator of soil moisture, which impacts processes such as plant growth. In addition to soil variables, the sensors capture variables like air temperature (°C), humidity (%), air pressure (mb), and rainfall data (mm). Battery voltage (vDC) is tracked

across all sensors to ensure consistent operation.

Hygro sensors measure humidity in the soil, and they often work using resistive methods to detect the moisture level of the soil based on the conductivity of the soil. As for the humidity, it uses materials that change in electrical properties and convert them into humidity readings. SEPA has 22 Hygro sensors measuring volumetric water content (VWC) directly, which provides valuable data but is insufficient for detailed spatial analysis. However, 9 DROPLET sensors measure other variables highly related to soil moisture, such as air temperature, rainfall, and air humidity, which can provide additional information on soil moisture patterns. The DROPLET sensors function as bucket rain gauges, which collect a certain amount of rain and record it as a measurement (Scottish Environment Protection Agency, 2025). These measurements can contribute to VWC predictions generated across the Elliot water catchment in any location, providing better spatial coverage of soil moisture patterns. The data dashboard is available at SEPA SensorNet IoT Portal.

The data supplied by SEPA supports environmental monitoring, which contributes to understanding soil dynamics in real time. The soil moisture properties are summarised in Table 1.1, and a more detailed explanation of these soil moisture indices can be found on the website (Scottish Environment Protection Agency, 2024). Figure 1.1 shows the distribution of the sensors in the study catchment, Elliot Water, and all the sensors are located alongside the river.

Table 1.1: Soil Moisture Variables

	Variables	Units
HYGRO	VWC	%
	Soil temperature	°C
	Conductivity	dS/m
	Air temperature	°C
	Humidity	%
	Battery voltage	vDC
DROPLET	Air temperature	°C
	Air pressure	mb
	Air humidity	%
	Rainfall	mm
	Battery voltage	vDC

Figure 1.1: Left: Elliot Water within Scotland. Right: SEPA sensor map. The blue icons represent the Droplet sensors, the green icons represent the Hygro sensors, and the red icons represent the locations having both Hygro and Droplet sensors.

1.2.2 COSMOS data

COSMOS-UK is a network of 47 sites across the UK that use cosmic-ray sensors to measure soil moisture, covering about 12 hectares per site. The data support farming, water management, flood forecasting, and land modelling. Managed by the UK Centre for Ecology & Hydrology and funded by the Natural Environment Research Council, COSMOS-UK has maintained well-managed VWC sensors since 2015. These sensors serve as the benchmark for VWC measurement (UK Centre for Ecology and Hydrology, 2024), with services maintained to the best possible standards. Modelling the relationship between VWC, air temperature, and precipitation in Balruddery (the nearest COSMOS sensor to Elliot Water) provides a better understanding of the potential relationships among these variables, and we can then apply this relationship to the sensors we have access to.

1.2.3 Satellite data

The satellite images from Copernicus describe the soil moisture of the soils topmost 5cm on a 1km (1°/112) spatial sampling (Copernicus Land Monitoring Service, 2024). It is derived from microwave radar data observed by the Sentinel-1 SAR satellite, which carries advanced radar to provide weather images of the Earth's surface (European Space Agency, 2025b). The satellite provides the soil water index (SWI), which is a key variable that provides an estimate of soil moisture conditions in the upper layers of the soil. The satellite has passed over the area of the Elliot Water every 3 or 4 days since 2015. Each satellite image covering Elliot Water includes 5 pixels horizontally by 19 pixels vertically. Such a collection of satellite images for the same region at different time points is called a Satellite Image Time Series (SITS). SITS can be seen as a stack of images or as a grid of time series: each pixel is associated with a time series. The SITS illustrates two views: the spatial view (a stack of pixels) and the temporal view (an array of time series).

Figure 1.2 shows a satellite image of the Elliot Water on 06/05/2021. Each cell within the grid

has a value for the cell location at the specific time point. The difference within the same grid represents the spatial variation within the area at the same time point. Figure 1.3 shows an example of the time series of pixel (3,13), showing how values change over time for a specific location. The satellite data visualisation combines both perspectives, which provide a comprehensive view of spatial and temporal patterns.

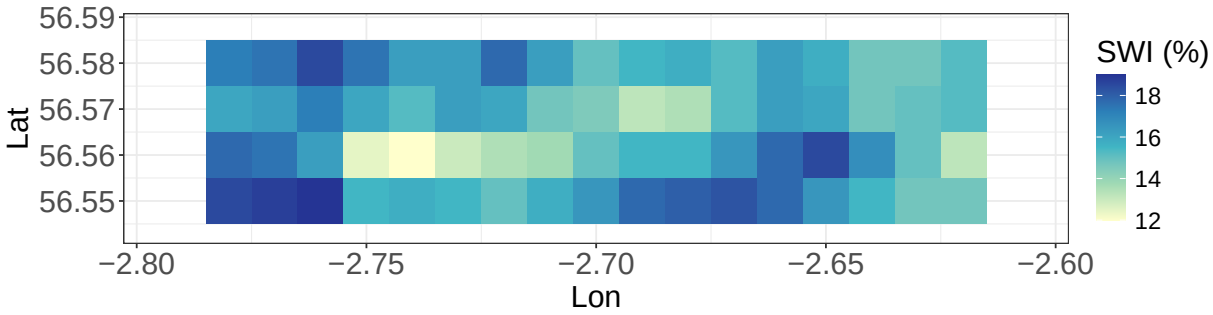


Figure 1.2: Scaled satellite-derived soil water index (SWI) over the Elliot Water catchment on 06/05/2021. SWI provides a relative measure of near-surface soil moisture, with lighter/darker shading indicating drier/wetter conditions across the catchment. This snapshot illustrates the spatial variability and gridded resolution of the satellite product that we later combine with in-situ sensor data in our data-fusion models.

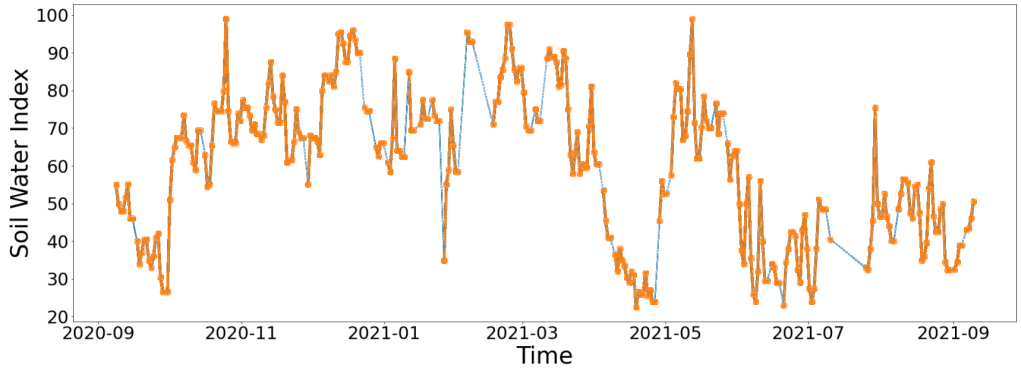


Figure 1.3: Time series of soil water index (SWI) from September 2020 to September 2021 for grid cell (3,13) in the Elliot Water catchment. Points indicate days with available SWI measurements, and blue dashed line segments highlight gaps between non-daily observations. The series shows both a clear seasonal pattern and substantial periods with missing data, motivating the need to handle irregular sampling and gaps carefully when constructing daily SWI covariates and when fusing SWI with in-situ soil moisture measurements in later chapters.

Unlike VWC, which measures the volume of water within a given volume of soil, SWI provides a relative measurement of soil moisture, typically scaled between 0 (dry) and 1 (saturated). SWI is often calculated using microwave-based remote sensing techniques, which are sensitive to the

dielectric properties of wet soil (Paciolla et al., 2020).

SWI and VWC share soil moisture dynamics across time. While VWC provides a direct measurement at a specific point location, SWI provides an indirect, comprehensive measurement, accounting for large-scale monitoring and modelling. By calibrating SWI with in-situ VWC measurements, it is possible to enhance the reliability of soil moisture estimates, bridging the gap between point data and gridded data. This integration supports a lot of applications, including water resource management and agricultural planning.

1.2.4 Elevation and soil type data

Elevation plays an important role in soil moisture dynamics as it influences precipitation, temperature, and hydrological processes, which in turn impact soil moisture. Higher elevations often get more rainfall and have lower temperatures, which reduces evapotranspiration and leads to soils retaining more moisture. In addition, elevation has impacts on vegetation and soil properties, which contribute to variations in soil moisture. Understanding these relationships is crucial for modelling soil moisture. The elevation data were obtained from the Open Elevation API (Open-Elevation, 2023), and the script to get the elevation data was developed by Andrew Elliott (Elliott, n.d.) with a resolution of 250m. Figure 1.4 shows the elevation of the study catchment, which displays the descending trend from northwest to southeast.

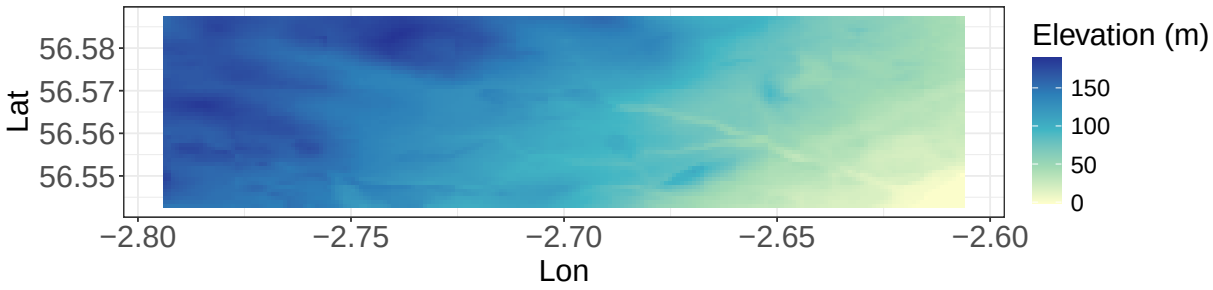


Figure 1.4: Elevation data visualisation for study catchment on a 250m resolution (Elliot Water)

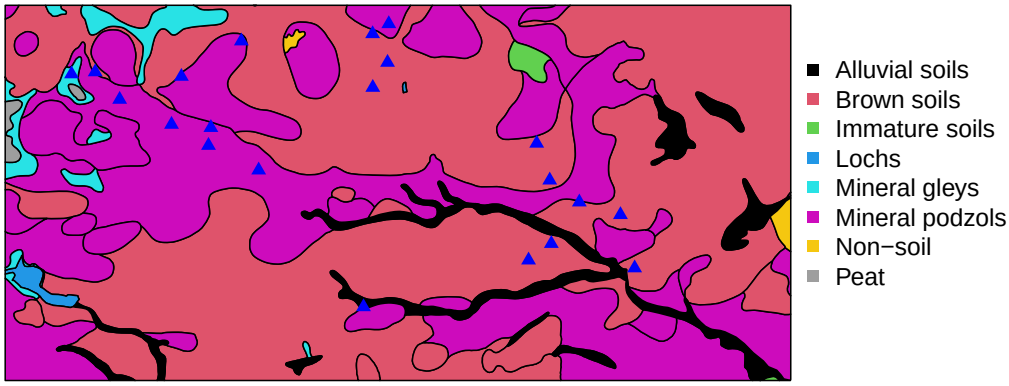


Figure 1.5: Soil type map for Elliot Water (James Hutton Institute, 2024)

The soil type data are obtained from the James Hutton Institute (James Hutton Institute, 2024). Figure 1.5 shows the soil type distribution in Elliot Water with SEPA sensors on top of it, which is dominated by the brown soils and mineral pozols represented by red and pink individually. The alluvial soils (black area) are where the river is located. Most of the sensors are in the mineral podzols soil type, one sensor is located in the mineral gleys, and the rest are in the brown soils. Table 1.2 shows the details about the legend:

Table 1.2: Descriptions of different soil types in the Elliot Water catchment.

Soil Type	Description
Alluvial soils	Associated with river valleys or floodplains, where soils are deposited by water.
Brown soils	Generally fertile and well-drained areas.
Immature soils	Soils that havent developed full horizons, often found in areas with recent geological activity.
Lochs	Lakes or water bodies.
Mineral gleys	Waterlogged soils due to poor drainage.
Non-soil	Surfaces not covered by soil, like bare rock or urban areas.
Peat	Rich in organic material, forming in waterlogged conditions, often in bogs.

Figure 1.6 shows the soil moisture level of cells within the satellite grid images grouped by soil type. The lochs and mineral gleys types each consist of only a single value, which shows no variability. Among the other three soil types, alluvial soil has a median of around 48% with moderate variation. Brown soil has a median of around 48% and shows greater variability than alluvial soil. Mineral podzols have a median of around 49%, showing the highest variability among these three soil types.

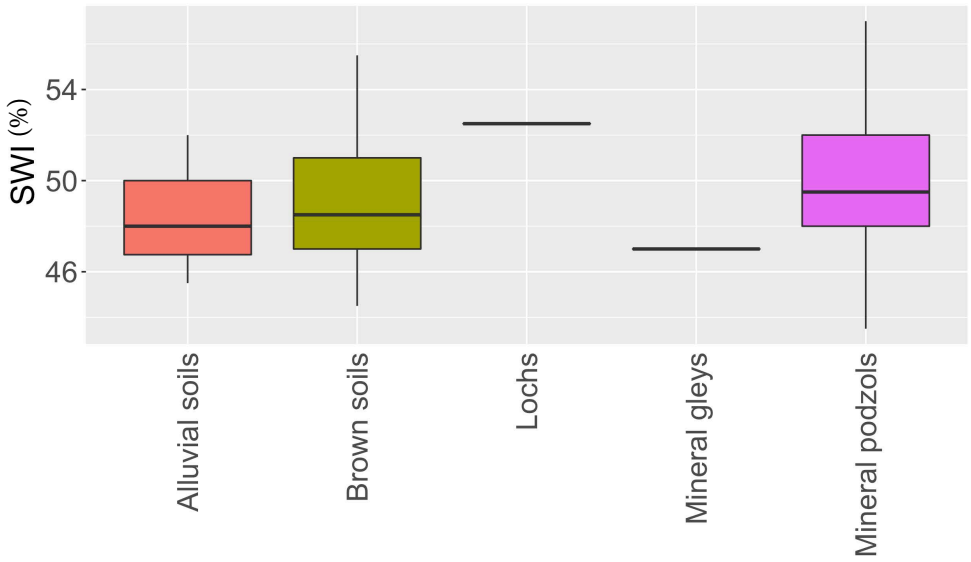


Figure 1.6: Boxplots of soil water index (SWI) by dominant soil type for grid cells in the Elliot Water catchment. Each box summarises the distribution of SWI across all grid cells of that soil type. For Lochs and Mineral gleys, there are very few grid cells, so the interquartile range collapses and the boxplot reduces to a single line, indicating that these categories should be interpreted with caution.

Overall, these considerations motivate a spatio-temporal data fusion method. Our aim is to assess whether combining in-situ point measurements with satellite gridded observations can deliver a more informative soil moisture map at a finer resolution, while providing uncertainty quantification. The following section summarises related work and highlights the gaps that motivate the data fusion framework.

1.3 Literature review

Spatio-temporal data fusion is a method to combine multiple datasets at different scales and resolutions to obtain insights into fields like environmental monitoring. This literature review summarises methodological advances and practical applications across three main data fusion frameworks: Bayesian hierarchical models, which offer statistically rigorous, uncertainty-aware fusion through techniques such as INLA-SPDE. Machine learning approaches, which apply classical algorithms and deep-learning architectures to detect patterns in high-dimensional data. This section also examines cross-cutting methods and validation strategies, and emerging data fusion trends, to highlight unresolved challenges and to suggest future research directions. Together, this review demonstrates the evolving trends in spatio-temporal fusion and outlines the key methodological trade-offs and application requirements.

1.3.1 Bayesian hierarchical models for spatial data fusion of point-referenced and gridded data

Spatial data fusion is a critical challenge in environmental and ecological statistics, where data are often from different sources, such as point measurements (e.g., sensors) and areal or gridded data (e.g., satellite pixels, administrative units). As verified for soil moisture in Section 1.2, these different data sources are often misaligned in their own spatial support and temporal frequency and have different measurement errors, which complicates the inference and prediction of the underlying spatial processes. Over the past two decades, Bayesian hierarchical models (BHM), especially those built upon latent Gaussian processes and the stochastic partial differential equation (SPDE) approach within the integrated nested Laplace approximation (INLA) framework, have become the method of choice for addressing these challenges. This section traces the evolution, key methodological developments, and applications, with special attention to change-of-support, misalignment, observation errors, and prediction at unobserved locations.

1.3.1.1 Change of support foundations

Fundamental work on spatial data fusion and the change-of-support problem began with BHM frameworks, which are known as Bayesian melding. This model merges point and gridded data through a latent spatial field and introduces aggregation methods to map the latent process to the specific spatial support and accommodate the support-specific measurement errors. For example, Gelfand et al. (2001) introduce a unified Bayesian approach for prediction across multiple combinations of point and block supports, applying fully Bayesian kriging and explicitly considering support aggregation and disaggregation in the likelihood. Gotway and Young (2002) point out the key problems when merging spatial datasets at different supports, resolutions, and locations, which are the change of support and the modifiable areal unit problem (MAUP). Fuentes and Raftery (2005) develop the Bayesian melding approach for fusing point observations and numerical model outputs, focusing on spatial environmental processes such as air pollution. The method assumes that both data types are from a shared latent ground truth Gaussian process, with explicit bias correction and uncertainty propagation, allowing improved spatial prediction and model evaluation. Wikle and Berliner (2005) propose a hierarchical conditioning approach to combine information across spatial scales, suitable for cases when inference is required only at specified resolutions. Gotway and Young (2007) extend these ideas by presenting a geostatistical framework compatible with GIS implementations, which explicitly incorporates data supports, handles covariate misalignment, and enables prediction with measures of uncertainty. Sahu et al. (2010) extend these models to the spatio-temporal domain to analyse point-referenced wet deposition data and gridded numerical model outputs within a fully Bayesian hierarchical context, explicitly addressing temporal aggregation and change-of-support issues. Nguyen et al. (2012) introduce Spatial Statistical Data Fusion (SSDF), a method for fusing massive spatial datasets observed at different supports. It deals with the change-of-support problem and achieves

scalability by using Fixed Rank Kriging (FRK) within a Spatial Random Effects (SRE) model, which provides a low-rank covariance representation for efficient computation.

Wilkie et al. (2019) address the problem of combining environmental measurements at mismatched spatial and temporal scales by developing a nonparametric statistical downscaling method that treats both in-situ and satellite observations as smooth functions over time, and links them through a Bayesian hierarchical model with spatially varying coefficients. By fitting basis function curves (e.g., a Fourier basis) to point-time and grid-cell-scale log(chlorophyll-a) data from Lake Balaton, the model generates predictions at any location and time points, which are complete with uncertainty estimates, without requiring temporal aggregation of the in-situ data. In a leave-one-out cross-validation, this approach outperforms purely spatial downscaling in accuracy (lower RMSE and MAE) and CIs coverage, demonstrating its effectiveness for fused spatio-temporal interpolation and its potential as a general tool for environmental data fusion. Wang and Furrer (2019) introduce the generalised spatial fusion model, which unifies latent Gaussian field approaches to point and area data fusion using flexible analytical tools such as low-rank approximations and explicit change-of-support matrices. This model framework has impacts on later computational approaches. Their following works include high-efficiency implementations using nearest neighbour Gaussian processes (NNGP) and comparisons with other fusion methods. Godoy et al. (2022) propose the Hausdorff-Gaussian process (HGP), which employs the Hausdorff distance to model spatial dependence in both point and areal data, and is competitive with other fusion models. Zhou and Bradley (2024) address multivariate and multiscale spatial data fusion from a Bayesian framework, proposing bivariate spatial models capable of handling measurements at different supports, thus extending established univariate methodologies. Cowles et al. (2009) present MCMC for efficient Bayesian estimation and prediction in hierarchical Gaussian models fusing point and areal data.

1.3.1.2 INLA-SPDE based BHM fusion models

Building upon this foundation, BHM, many recent studies have obtained advanced computational efficiency and flexibility through using INLA-SPDE methods that represent the latent field via Gaussian Markov random fields on triangulated meshes. This approach enables modelling of change-of-support problems, allowing for integration of both point and gridded data sources.

Moraga et al. (2017) introduce a geostatistical model for point-level and area-level spatial data within the INLA-SPDE framework. This pioneering work involves computing basis function integrals over grid supports, allowing both point data and grid data to contribute to the inference of a latent Gaussian field. Many later studies about point data and grid data fusion reference and build upon this foundational work. Wilson and Wakefield (2020) use SPDE-based continuous spatial models to introduce area-level random effects via a latent Gaussian field, while also incorporating point-level data. They then explore different computational strategies for Bayesian

inference, which apply INLA to linear models and turn to fully Bayesian Hamiltonian Monte Carlo or empirical Bayes techniques for more complex or nonlinear relationship settings.

Recent literature extends these foundational models to address modern complexities. For example, Zhong et al. (2025) address preferential sampling in joint models, modifying the INLA-SPDE fusion approach to jointly model the process generating monitoring sites and the spatial latent field, thereby correcting potential bias from non-random location selection. Roksvåg et al. (2021) develop multifield geostatistical models combining point and nested areal observations (such as precipitation) within an INLA-SPDE framework. The two-field models decompose climatic and yearly spatial effects, providing detailed insights into spatial variability. He and Wong (2024), and Villejo et al. (2023), apply INLA-SPDE data fusion to spatio-temporal settings, integrating in-situ and remote sensing observations as well as multiple likelihood components to improve prediction over space and time. Chacón-Montalván et al. (2024) propose an INLA-SPDE model explicitly handling change-of-support not only for responses but also for covariates, modelling both as latent Gaussian processes with potentially rectilinear supports, which enables propagation of uncertainty due to misalignment in predictors. Suen et al. (2025) introduce a Bayesian disaggregation framework for spatially misaligned data, applying an iteratively linearised integration method via INLA. This framework supports different scenarios for covariate raster at full resolution, aggregation, and point values, propagating uncertainty when covariate information is incomplete. Zhong and Moraga (2023) compare Bayesian melding and downscaler approaches within the INLA-SPDE framework, presenting the model performance in fusing spatially misaligned data.

Cameletti et al. (2019) and Forlani et al. (2020) further extend the INLA-SPDE fusion framework to health and air pollution applications, modelling both point and gridded data sources such as monitoring measurements and model gridded outputs, while accounting for spatial and temporal misalignment and different sources of uncertainty. These models allow joint prediction of pollutant concentrations and demonstrate improved predictive performance by combining information from multiple data sources. Villejo et al. (2025) propose a data fusion model for meteorological variables, addressing sparse observational coverage by incorporating numerical forecast models as an extra data source. This approach models both classical error structures for point observations and structured additive/multiplicative biases for gridded forecasts within the INLA-SPDE framework, evaluated through cross-validation and simulation. Roksvåg et al. (2021) fuse areal runoff and point precipitation in annual runoff prediction, demonstrating the advantage of the joint model. Validation across these applications involves leave-one-out cross-validation (LOOCV) and usage of proper scoring rules, with consistent evidence that Bayesian fusion models outperform single-source models or non-hierarchical alternatives.

1.3.1.3 Limitations and key challenges

The literature review of Bayesian Hierarchical Models (BHMs) highlights several common methods and challenges, and demonstrates a mature field for Bayesian spatial data fusion of point and gridded data. For the change-of-support and misalignment issues in INLA-SPDE based fusion frameworks, change-of-support and misalignment issues are commonly addressed through explicit aggregation operators. For the measurement error, most methods break down the measurement error into two parts: measurement errors at individual stations and aggregation errors from averaging over areas. While theoretical frameworks for multivariate and multi-scale data fusion models exist, their application to large scale real world problems remains very limited. Covariate misalignment and uncertainty, particularly in spatially misaligned predictors, represent an active area of methodological extension. Computationally, INLA dominates implementations for latent Gaussian models due to its efficiency and flexibility with direct modelling of support and misalignment at both the observation and covariate level. Though MCMC or hybrid algorithms are preferred for non-Gaussian or highly complex model structures. In addition, model validation practices, including cross-validation and scoring rules, are widely adopted to ensure the reliability of fused predictions. As for the real data application, there are real data application studies across environmental, meteorological, and hydrological fields that consistently validate these approaches, with ongoing research focused on addressing computational scalability, misaligned covariates, and rich forms of data misalignment. These topics demonstrate both the maturity and evolving nature of BHM-based fusion frameworks.

1.3.2 Machine learning (ML) framework for the spatio-temporal data fusion

Traditional data fusion models based on a Bayesian hierarchical modelling (BHM) framework effectively capture spatial dependencies across multiple data levels and produce posterior uncertainty estimates, offering interpretable, probabilistic-based insights in research fields such as environmental monitoring and health sciences. However, BHMs usually need careful choice of priors and rely on computationally intensive sampling or variational methods, which become infeasible when dealing with very large datasets, high-dimensional features, or complex model structures. By contrast, modern machine learning (ML) approaches offer scalable, data-driven flexibility: methods like support vector machines, random forests, neural networks, and ensemble tree approaches can automatically learn hidden, nonlinear patterns from huge, heterogeneous data and benefit from efficient parallel optimisation. Although traditional ML models do not inherently quantify uncertainty, recent studies, such as distribution-free uncertainty frameworks applied to ensemble trees, including conformal prediction and quantile regression, are now able to provide not only point predictions but also formal confidence measurement. Consequently, ML-based fusion approaches can combine the interpretability and uncertainty measurement of BHMs while avoiding their computational and modelling limitations, making the way for

uncertainty-aware real data applications. Within this ML fusion framework, gradient-boosted decision trees outperform because of their rapid, additive learning and strong regularisation. Below, the review starts with a wide class of ML-based fusion methods and how they address uncertainty, then focuses specifically on gradient boosting implementations, such as XGBoost, for applications requiring uncertainty measurements.

1.3.2.1 Classical machine learning approaches

Early classical machine-learning methods mainly focused on combining physical knowledge with classical ML models. For example, the thermal-inertia theory, which links diurnal land-surface temperature (LST) amplitudes to SM, is combined with regression trees or support vector machines (SVM). Regression-tree models leveraging Moderate Resolution Imaging Spectroradiometer land-surface temperature (MODIS LST), vegetation indices (e.g., NDVI), seasonal indicators, and soil texture variables have achieved unbiased root-mean-square errors (ubRMSE) of 0.05-0.07 m^3 at 1 km resolution in southeastern Australia (Merlin et al., 2012). In addition, SVM models combined with spatial weighting achieved a correlation of 0.68 and RMSEs near 0.08 m^3 over Oklahoma (Kim et al., 2018). Self-regularising regression frameworks also show great performance in a temporal perspective by dynamically adapting to temporal changes, successfully tracking SM dynamics through a corn growing season in Texas (Hernández-Sánchez et al., 2019).

1.3.2.2 Deep learning architectures

More recently, Deep learning (DL) has outperformed classical methods by leveraging the fusion of multiple data sources. In recent work, deep learning uses a unified framework for merging heterogeneous remote-sensing and ground-based data into high-resolution maps. For example, Convolutional neural networks (CNNs) integrate inputs such as radar backscatter (Sentinel-1), microwave brightness temperatures (SMAP), optical data (Sentinel-2/MODIS), and other environmental covariates (e.g., terrain, soil texture) to learn joint features representing the data characteristics. Residual connections and spatial weighting layers are used to keep the inter-pixel heterogeneity, which helps the model achieve ubRMSE <0.05 m^3 (Li et al., 2023; Liu et al., 2020). In summary, these ML pipelines first use several convolutional layers to learn spatial patterns at different scales. Then, a final regression layer transforms those learned features into soil moisture values for each pixel. By training on ground measurements spread across the study area, the model learns to choose the best data source, optical data (e.g., Sentinel-2) under clear sky and switching to radar/microwave inputs (e.g., SMAP, Sentinel-1) when it's cloudy or heavily vegetated areas. The output product is a seamless soil moisture map at 30-320 m resolution that performs reliably across many land covers, including deserts, croplands, forests, and other land covers (Huang et al., 2022; Batchu et al., 2022).

While deep learning methods perform very well in merging soil moisture data, they have some major drawbacks. First, these models need tons of labelled training data, which is hard to get in areas with few ground sensors or constant cloud cover. This will probably cause the overfitting problem, meaning they work poorly where data is limited (Huang et al., 2022). Second, they are very computationally expensive because they require powerful computers and take hours on a multi-core workstation. This makes them hard to use in places with limited resources or for real-time tasks (Li et al., 2023). Third, it's hard to understand how they make the inference, which makes it challenging to find errors or check if the results make sense, especially when using different data types like radar or satellite images (Batchu et al., 2022). Lastly, these models often fail when used in new areas or seasons they weren't trained on, which means they have really poor generalisation. For example, a model trained on dry climates might perform very poorly in tropical regions (Ma et al., 2023). To fix these issues, researchers need to develop simpler models, use methods that require less labelled data, and add physics-based rules to make the models more reliable and adaptable.

1.3.2.3 Spatio-temporal cross validation

Spatio-temporal multi-source fusion generates very high temporal continuity and effectively resolves the gap problem caused by cloud cover or sensor limitation (Huang et al., 2022; Jing et al., 2024). For validation, a multi-site and space-time blocked cross-validation scheme is adopted, using both RMSE and correlation metrics, to avoid overfitting, particularly in heterogeneous terrain conditions (Huang et al., 2022; Mao et al., 2022; Wei et al., 2019).

However, non-separable spatio-temporal covariance modelling and a fully hierarchical Bayesian fusion framework remain very rare, and only a few multi-support, uncertainty propagating frameworks exist. Most machine learning fusion models, although they show high performance, still treat gridded values simply as covariates at point locations and have limited explicit modelling of support mismatch or different support error structure. In agricultural settings, it is also observed that topographic and vegetation variables often play more important roles than soil texture. Despite these advances, validation against sparse in-situ networks remains challenging, necessitating specialised airborne and field campaign data (Hernández-Sánchez et al., 2019). Future research must focus on model generalisation across climates and land covers, automatic feature selection or reducing dependence on optical inputs, and integration of in-situ observations to better assess spatial-pattern support (Senanayake et al., 2024).

1.3.3 Uncertainty in ML fusion

Accurate prediction with uncertainty quantification for unobserved locations is critical for many spatio-temporal data fusion models. This often requires the fusion of multiple data sources, such

as regularly gridded data and irregular point observations, and the application of machine learning methods capable of exploiting such fused features. Additionally, robust uncertainty quantification is essential, particularly under realistic scenarios exhibiting spatial and temporal dependence. The prediction without an uncertainty qualification is no different from a wild guess.

XGBoost, a gradient-boosted decision tree framework, is widely used in spatio-temporal prediction due to its ability to handle high-dimensional, heterogeneous input features. Its special application is to the fusion of spatio-temporal grid data and point observations for supervised learning at new locations or times. However, most literature independently discusses either data fusion or uncertainty quantification, and few studies directly address their integrated use in a dependency-aware model framework.

Conformal prediction is a model-agnostic framework for producing prediction intervals with finite-sample coverage guarantees. Standard conformal prediction relies on exchangeability assumptions, which are violated in spatio-temporal contexts due to spatial and temporal autocorrelation. Recent research has focused on adapting conformal prediction to respect spatial and temporal dependence, for instance, through localised calibration, clustering, or weighted nonconformity scores. Among the references from the previous study, only one paper, GeoConformal prediction (Lou et al., 2024a), explicitly fuses spatial grid features and point data using XGBoost and applies a geography-aware conformal calibration scheme. This approach addresses the integration of feature-level grid-point fusion with local dependency-weighted uncertainty quantification, focusing on spatial regression tasks such as housing price prediction.

Several papers Zhou et al. (2024); Lin et al. (2022); Mao et al. (2024); Hajibabaee et al. (2024) propose advancements in conformal prediction for spatial and temporal regression, introducing approaches such as localised quantile regression, block-based calibration, and data-dependent weighting of nonconformity scores. These studies focus on improving prediction interval estimation under dependence structures but generally remain agnostic to the choice of regression model and do not incorporate fusion of gridded and point data with XGBoost.

In summary, the literature provides strong support for dependence adapted conformal prediction methods and demonstrates the use of XGBoost in feature fusion for spatial and spatio-temporal prediction. The integration of both approaches, particularly for complete spatio-temporal fusion and dependence aware conformal uncertainty quantification using XGBoost, appears to be an unknown area, with GeoConformal representing a leading example for spatial dependence settings (Lou et al., 2024a).

1.4 Methodological background of basic spatial and temporal analysis

This section gives a high-level overview of the concepts and methods used throughout this thesis, with full methodology details presented in each chapters methodology section. It begins by outlining the fundamental concepts and approaches that support this work. Next, it outlines the background methodology, with spatial and temporal processes considered separately at first.

1.4.1 Spatial processes

Spatial autocorrelation refers to the pattern in which observations at nearby locations have more similar values than those further apart, violating the assumption of independent observations (Tobler, 1970; Cressie, 1993). The spatial dependence means that each new observation in a clustering area contributes less new information, which can bias the classical inference. For example, it will underestimate standard errors and inflate the Type I error rates if autocorrelation is ignored in the modelling (Dormann et al., 2007). However, modelling spatial autocorrelation can remedy this situation: geostatistical methods (e.g. kriging) use the covariance among neighbouring points to improve prediction accuracy and provide more realistic uncertainty estimates (Cressie, 1993; Wackernagel, 2003). For example, soil moisture often exhibits positive spatial autocorrelation in wet or dry clusters (spatial clustering) with a range of influence beyond which measurements become almost independent (Western and Blöschl, 1999). Recognising such spatial structure is crucial for reliable spatial predictions or upscaling of soil moisture data in environmental statistics (Cressie, 1993; Wackernagel, 2003).

A spatial stochastic process is a collection of random variables

$$\{Z(s) : s \in D \subset \mathbb{R}^d\},$$

where each $Z(s)$ represents the variable of interest (e.g. soil moisture) at location s (Journel and Huijbregts, 1976).

It distinguishes two stationarity assumptions:

- **Second order stationarity:**

$$\mathbb{E}[Z(\mathbf{s})] = \mu, \qquad \mathrm{Cov}\big(Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h})\big) = C(\mathbf{h}), \quad \mathbf{s}, \mathbf{h} \in \mathbb{R}^d,$$

  where the mean is constant and the covariance depends only on the lag vector h (Cressie, 1993).

- **Intrinsic stationarity:**

$$\mathbb{E}\big[Z(\mathbf{s}+\mathbf{h})-Z(\mathbf{s})\big]=0, \quad Var\big[Z(\mathbf{s}+\mathbf{h})-Z(\mathbf{s})\big]=2\,\gamma(\mathbf{h}),$$

  where $\gamma(\mathbf{h})$ is the semi-variogram. Intrinsic stationarity requires only that increments are stationary.

Variogram and covariance functions

In spatial statistics, stationarity means that the key characteristics of a spatial process do not change with location. The most common type is second-order stationarity, where the process has a constant mean and a covariance structure that depends only on the distance (or lag) between two points, not on their absolute locations (Cressie, 1993; Chilès and Delfiner, 2012). This assumption implies that the data are spatially homogeneous and that the semivariogram depends only on the lag h. While useful for modelling, this assumption is often too strong in real data applications.

A more flexible assumption is intrinsic stationarity, which does not require a full covariance function. Instead, it assumes that the difference between values at two locations, $Z(s+h)-Z(s)$, has a mean of zero and a variance that depends only on h. This variance is called the variogram, and it describes how variability increases with distance.

When a spatial process is non-stationary, its mean or variance changes across space. This can happen when there is a trend (e.g., values increase from north to south) or when different areas have different levels of variability. In such cases, the variogram may continue to increase with distance and never level off, making it hard to model the true spatial structure. If these patterns are not corrected, predictions can become biased and uncertainty estimates unreliable (Cressie, 1993).

To address this, spatial analysts use several diagnostic tools. They include visualisation, empirical variograms, and regressions on spatial covariates to detect trends (Banerjee et al., 2003). If non-stationarity is detected, detrending methods are applied. A common approach is to fit a trend model (e.g., regression on coordinates or elevation) and subtract it, leaving residuals that are almost stationary (Chilès and Delfiner, 2012). Another approach is to use a moving window or divide the area into subregions and remove the local mean. Once the trend is removed, standard geostatistical tools, such as kriging and variogram modelling, can be used.

Stationarity Assumptions

The semivariogram is defined by

$$2\,\gamma(h) \;=\; Var\big[Z(s+h)-Z(s)\big] \;=\; \mathbb{E}\big[(Z(s+h)-Z(s))^2\big].$$

Under second-order stationarity, it shows

$$\gamma(h) \;=\; \sigma^2 - C(h), \quad C(h) = Cov\big(Z(s), Z(s+h)\big), \quad \sigma^2 = C(0).$$

The empirical (methodofmoments) variogram estimator is

$$\hat{\gamma}(h) = \frac{1}{2\,|N(h)|} \sum_{(i,j)\in N(h)} \big[Z(s_i) - Z(s_j)\big]^2,$$

where $N(h) = \{(i,j)\colon \|s_i - s_j\| \approx h\}$ (Journel and Huijbregts, 1976). The variogram $\gamma(h)$ increases with $||h||$ and typically levels off at a sill value equal to the process variance when h exceeds the range of spatial correlation (Cressie, 1993).

It estimates an empirical variogram from the data and then fits a standard variogram model to it (e.g. spherical, exponential, or Matérn models) following guidelines in (Chilès and Delfiner, 2012). In particular, the Matérn covariance function is often chosen for its flexibility. It includes a tunable smoothness parameter ν that controls the differentiability of the field, making it capable of representing various smoothness levels of spatial processes (Rasmussen and Williams, 2006).

Common theoretical models are defined as follows:

$$\begin{aligned}
\text{Spherical:} \quad & \gamma(h) = \begin{cases} c_0 + c_s\Big[1.5\frac{h}{a} - 0.5\big(\frac{h}{a}\big)^3\Big], & 0 < h \le a, \\ c_0 + c_s, & h > a, \end{cases} \\
\text{Exponential:} \quad & \gamma(h) = c_0 + c_s\Big[1 - e^{-h/a}\Big], \\
\text{Matérn:} \quad & \gamma(h) = c_0 + c_s\Big[1 - \tfrac{1}{2^{\nu-1}\Gamma(\nu)}\big(\tfrac{h}{a}\big)^{\nu} K_\nu\big(\tfrac{h}{a}\big)\Big],
\end{aligned}$$

where c_0 is the nugget, $c_0 + c_s$ the sill, and a the range.
Nonstationarity To remove largescale trends, fit a drift surface $m(s)$ (e.g. polynomial or regression on covariates) and analyse residuals

$$Z_{\text{res}}(s) \;=\; Z(s) - m(s).$$

Alternatively, compute local variograms in moving windows to detect spatially varying correlation.

1.4.2 Kriging

1.4.2.1 Simple kriging (SK)

Simple kriging assumes that the process mean μ is known and constant over the domain (Journel and Huijbregts, 1976). The SK predictor at a target location s_0 is

$$\hat{Z}(s_0) = \mu + \sum_{i=1}^{n} \lambda_i \big[Z(s_i) - \mu \big],$$

where the weights $\{\lambda_i\}$ are chosen to minimise the mean squared error

$$\mathrm{MSE} = Var\big[\hat{Z}(s_0) - Z(s_0)\big] = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \, \sigma_{ij} \; - \; 2\sum_{i=1}^{n} \lambda_i \, \sigma_{i0} \; + \; \sigma_{00}.$$

Differentiating with respect to λ_i and setting to zero yields the linear system

$$\sum_{j=1}^{n} \lambda_j \, \sigma_{ij} \;=\; \sigma_{i0}, \quad i = 1, \dots, n,$$

or in matrix form $\Sigma \lambda = \sigma_0$, where $\Sigma = [\sigma_{ij}]$ is the $n \times n$ covariance matrix among the samples and $\sigma_0 = (\sigma_{10}, \dots, \sigma_{n0})^{\top}$ is the covariance vector with $Z(s_0)$. The minimised MSE, known as the kriging variance, is

$$\sigma^2_{\mathrm{SK}} = \sigma_{00} - \sigma_0^{\top} \lambda .$$

1.4.2.2 Ordinary kriging (OK)

Ordinary kriging is a best linear unbiased estimator (BLUE) for a second-order stationary random field $Z(s)$ with unknown constant mean μ (Cressie, 1993; Chilès and Delfiner, 2012). It seeks a predictor

$$\hat{Z}(s_0) \;=\; \sum_{i=1}^{n} \lambda_i \, Z(s_i),$$

that minimises the mean squared prediction error

$$Var\big[\hat{Z}(s_0) - Z(s_0)\big] \quad \text{subject to} \quad \mathbb{E}\big[\hat{Z}(s_0) - Z(s_0)\big] = 0.$$

The unbiasedness constraint $\sum_i \lambda_i = 1$ together with Lagrange multiplier ν leads to the kriging system (Journel and Huijbregts, 1976; Wackernagel, 2003):

$$\begin{pmatrix} C(h_{11}) & \cdots & C(h_{1n}) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C(h_{n1}) & \cdots & C(h_{nn}) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \nu \end{pmatrix} = \begin{pmatrix} C(h_{10}) \\ \vdots \\ C(h_{n0}) \\ 1 \end{pmatrix},$$

where $C(h) = \sigma^2 - \gamma(h)$ is the covariance model derived from the variogram, $h_{ij} = \|s_i - s_j\|$. The resulting kriging variance is

$$\sigma_K^2(s_0) = C(0) - \sum_{i=1}^{n} \lambda_i C(h_{i0}) - \nu.$$

Unbiasedness and minimum variance Properties
By construction, OK satisfies

$$\mathbb{E}[\hat{Z}(s_0)] = \mu \quad \text{and} \quad Var[\hat{Z}(s_0) - Z(s_0)] \leq Var[\hat{Z}'(s_0) - Z(s_0)]$$

for any other linear unbiased estimator $\hat{Z}'$ (Chilès and Delfiner, 2012).

1.4.2.3 Universal kriging (UK)

Universal kriging (UK) allows a deterministic trend $m(s)$ in the mean, which denotes the large-scale pattern and a residual term $\varepsilon(s)$ denoting the small-scale spatial variation:

$$Z(s) = m(s) + \varepsilon(s), \quad m(s) = \sum_{k=0}^{p} \beta_k f_k(s),$$

where $\{f_k(s)\}$ are known covariates (e.g. elevation) and $\varepsilon(s)$ is a zero-mean stationary residual (Journel and Huijbregts, 1976; Cressie, 1993). The trend model can be expressed as a linear combination of basis functions. The UK system adds constraints

$$\sum_{i=1}^{n} \lambda_i f_k(s_i) = f_k(s_0) \quad (k = 0, \ldots, p),$$

ensuring the estimator $\hat{Z}(\mathbf{s}_0)$ preserves the trend structure: $\mathbb{E}[\hat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)] = 0$. Drift parameters β_k can be estimated jointly in the kriging system or by regression followed by kriging of residuals (Wackernagel, 2003).

UK system simultaneously estimates β_k and spatial weights via extended kriging equations:

$$\begin{bmatrix} \boldsymbol{\Gamma} & F \\ F^\top & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\nu} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\gamma}_0 \\ f_0 \end{bmatrix}$$

solves for both the optimal kriging weights $\boldsymbol{\lambda} \in \mathbb{R}^n$ and Lagrange multipliers $\boldsymbol{\nu} \in \mathbb{R}^{(p+1)}$, which enforce the unbiasedness constraints $\sum_{i=1}^{n} \lambda_i f_k(\mathbf{s}_i) = f_k(\mathbf{s}_0)$ required for trend reproduction. This dual solution ensures spatial predictions $\hat{Z}(\mathbf{s}_0) = \boldsymbol{\lambda}^\top \mathbf{Z}$ maintain both minimum variance and structural consistency with the specified trend model.

1.4.2.4 Block and area-to-point kriging

Block kriging predicts the average over an area B:

$$Z(B) = \frac{1}{|B|}\int_B Z(u)\,du, \quad C(h_{iB}) = \frac{1}{|B|}\int_B C(\|s_i - u\|)\,du,$$

and replaces point covariances with block-to-point covariances in the kriging system (Chilès and Delfiner, 2012). Area-to-point kriging further solves for pointwise estimates consistent with those block averages (Cressie, 1993).

1.4.2.5 Kriging Diagnostics

Cross-validation is used to assess performance (Cressie, 1993; Wackernagel, 2003):

- **Mean Error (ME):** $\frac{1}{n}\sum_i[\hat{Z}_{-i}(s_i) - Z(s_i)]$.
- **Root-Mean-Square Error (RMSE):** $\sqrt{\frac{1}{n}\sum_i[\hat{Z}_{-i}(s_i) - Z(s_i)]^2}$.
- **Standardised Errors:** $e_i^* = \frac{\hat{Z}_{-i}(s_i) - Z(s_i)}{\sigma_K(s_i)}$, with $\text{RMSSE} = \sqrt{\frac{1}{n}\sum_i (e_i^*)^2} \approx 1$.
- **Variance-to-Error Ratio:** Compare mean kriging variance to MSPE; a ratio near 1 indicates well-calibrated uncertainty.

1.4.2.6 Implementation Steps

1. Compute empirical variogram $\hat{\gamma}(h)$ via method-of-moments (Cressie, 1993).
2. Fit a theoretical variogram model (spherical, exponential, Matérn) to $\hat{\gamma}(h)$ (Chilès and Delfiner, 2012).
3. Build covariance matrix $C(h_{ij}) = \sigma^2 - \gamma(h_{ij})$ and cross-covariance vector $C(h_{i0})$.
4. Solve the kriging system for $\{\lambda_i\}$ and ν.
5. Compute predictions $\hat{Z}(s_0)$ and variances $\sigma_K^2(s_0)$.

1.4.3 Temporal differencing

To handle non-stationary trends, we differentiate the time series:

$$\nabla_t Y(s,t) \equiv Y(s,t) - Y(s,t-1).$$

This removes slow drift and helps stabilise the mean. When there is clear seasonality, we also use a seasonal difference $\nabla_T Y(s,t) = Y(s,t) - Y(s,t-T)$. We check the ACF and PACF after differencing to confirm stationarity.

1.4.4 Augmented Dickey-Fuller (ADF)

The augmented Dickey-Fuller (ADF) test evaluates stationarity by testing for a unit root in a time series (Dickey and Fuller, 1979). The definition of ADF is as follows:

$$\text{ADF Statistic} = \frac{\hat{\gamma}}{\text{SE}(\hat{\gamma})}, \tag{1.1}$$

where $\hat{\gamma}$ is the estimated coefficient of the lagged dependent variable and $\text{SE}(\hat{\gamma})$ is the standard error of $\hat{\gamma}$.

1.4.5 Separable and non-separable spatio-temporal covariance

Let $Z(\mathbf{s},t)$ be second-order stationary in time and isotropic in space, with $C(h,u) = \text{Cov}\{Z(\mathbf{s},t), Z(\mathbf{s}',t')\}$, $h = \|\mathbf{s}-\mathbf{s}'\|$, $u = |t-t'|$.

Separable covariance

A covariance is separable if

$$C(h,u) = \sigma^2 C_S(h) C_T(u), \qquad C_S(0) = C_T(0) = 1,$$

e.g., Matérn-in-space $\times$ AR(1) in time. This implies no spacetime interaction and yields a Kronecker structure for computations.

Non-separable (mixtures of separable components)

Broad non-separable classes arise by mixing valid separable components:

$$C(h,u) = \int C_S\big(h;\xi\big)\, C_T\big(u;\xi\big)\, \mathrm{d}F(\xi),$$

where for each ξ the pair $\{C_S(\cdot;\xi), C_T(\cdot;\xi)\}$ is valid with $C_S(0;\xi) = C_T(0;\xi) = 1$, and F is a non-negative mixing measure. Unless F is degenerate or either factor is ξ-constant, C is non-separable (De Iaco et al., 2002; Ma, 2002, 2003).

1.5 Research gaps and objectives

Despite progress in data fusion methods, several methodological and application specific challenges still exist. First, differences in spatial and temporal resolution (change of support) between sensors and satellite data are not always explicitly addressed, which can impact prediction performance. Second, many machine learning models treat spatial data as independent inputs, ignoring spatial and temporal dependencies. Third, uncertainty quantification is often missing, making it hard to assess the reliability of predictions. This thesis aims to address these gaps by developing and comparing fusion methods that combine in-situ and satellite soil moisture data, using both a geostatistical INLA-SPDE framework (primary focus) and a machine learning method. The

research objectives are: to build a full data fusion framework with INLA-SPDE that jointly combines point (in-situ) and gridded satellite data, explicitly handles change-of-support and covariate misalignment, and produces high-resolution soil moisture maps with posterior uncertainty, and to develop a machine learning fusion method (XGBoost) equipped with calibrated uncertainty via conformal prediction.

1.6 Thesis structure

This thesis is organised as follows. Chapter 1 describes the datasets used in this thesis, including in-situ sensor networks and satellite data and introduces the basic methodological methods and concepts of spatial modelling and time series analysis. Chapter 2 presents an exploratory analysis of their spatial and temporal properties and uses geostatistical methods to explore the spatial and temporal patterns within the study catchment. Chapter 3 presents the issue of misaligned covariates in spatial regression. Chapter 4 focuses on the spatio-only INLA-SPDE data fusion model. Chapter 5 develops and evaluates a patio-temporal INLA-SPDE data fusion model. Chapter 6 presents machine learning approaches for soil moisture data fusion. Chapter 7 concludes the thesis and outlines some future directions.

Specifically, the contributions of this thesis are: a spatio-temporal regression with spatially misaligned covariates, the development of a spatio-temporal INLA-SPDE data-fusion model to combine sensor and satellite data. And a machine learning fusion method that encodes spatial structure in XGBoost (via a Laplacian-penalised loss) and combines uncertainty quantification through spatio-temporal conformal prediction, which explores the classical data fusion approaches with modern ML.

Chapter 2

Exploratory Data Analysis of Soil Moisture Data from In-Situ Measurements, COSMOS, and Satellite Observations

Chapter 1 provides an introduction to the research background, a description of the specific river catchment of interest to the study, a description of each data source, a comprehensive literature review of existing data fusion methods, and the methodology, including several common approaches used throughout this thesis. Chapter 2 provides an exploratory analysis of in-situ sensors and satellite images to get a comprehensive overview of the real data and guide further exploration and study.

The study focuses on the catchment area of Elliot Water, using multiple data sources to uncover spatial-temporal patterns within the data, which includes data from Scottish Environment Protection Agency (SEPA) environmental monitoring sensors (Section 1.2.1), COSMOS soil moisture sensors (Section 1.2.2), and Copernicus satellite images (Section 1.2.3). The multiple datasets provide a multi-scale and multi-source perspective, which provides a robust exploration of spatial-temporal patterns within the study catchment. This primary exploration not only uncovers the underlying features within the dataset but also inspires further studies.

There are three soil moisture data sources of interest in this study: SEPA data, COSMOS data and Copernicus data. They provide the soil moisture data for Elliot Water (Figure 1.1) at different spatial and temporal resolutions. The SEPA volumetric water content (VWC) data is recorded every 15 minutes, and COSMOS and Copernicus data are available daily. As for the preprocessing steps, the cleaning procedures will be discussed, such as how to handle the missing values, remove outliers, and correct errors within the data. The transformation or standardisation done to the datasets will be explained in this section. In addition, the data are misaligned, so the alignment of the spatial-temporal data will be discussed.

In Section 2.1, for the SEPA data, the focus is on examining the relationships among different variables. In Section 2.2, time decomposition is employed in the COSMOS data to investigate the temporal trends in soil moisture, as the COSMOS monitoring network provides well-maintained near-real-time soil moisture data, so it serves as a benchmark for soil moisture measurements in the UK. In Section 2.3, given the large spatial coverage of satellite images, Linear Models (LM) and Generalised Additive Models (GAM) are fitted to the SWI data to analyse the spatial patterns within the satellite data. Section 2.4 will investigate the relationship between VWC and SWI to uncover their alignment, which will be crucial for the data fusion method discussed in Chapter 4.

2.1 SEPA data

Sensor data began in 08/09/2020. The air temperature and soil temperature show an artefact due to the integer overflow, which happens when an arithmetic operation on integers produces a numeric value that exceeds the range that can be represented by the given number of digits, either above the maximum or below the minimum limit (Wikipedia contributors, 2025). Due to integer overflow, where calculations exceed the maximum storable value, the sensor's retrieval algorithm produced extremely high readings when attempting to process negative values. The VWC is the percentage water content of the soil, which is generally at a minimum during the summer and at a maximum in winter (Quinn et al., 2020). Soil temperature is also an important measurement of soil moisture because previous studies show that an increase in moisture content decreases the soil temperature differences between daytime and nighttime. This may give a hint to farmers and scientists that protecting the plant root system against sharp and sudden changes in soil temperature will be beneficial (Al-Kayssi et al., 1990). During the COVID-19 lockdown, fieldwork was not allowed, and the sensors used are prototype sensors still in the first stage of development and testing. Therefore, the sensors are saturated and have not been cleaned during the COVID-19 lockdown period. In the humidity readings, this appears as repeated 100% humidity readings, which reflect sensor saturation rather than true measurements. However, sensors that did not have this issue will help to adjust the data.

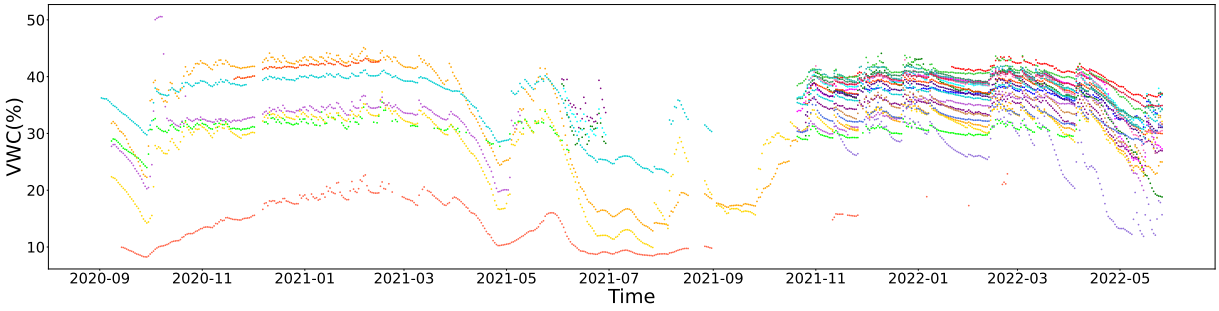


Figure 2.1: Time series of 15-minute volumetric water content (VWC) recorded by SEPA sensors in the Elliot Water catchment. Each coloured line is a different sensor. The plot shows higher VWC in winter, lower VWC in summer. Some sensors are generally wetter or drier than others, and there are short gaps where sensors were not recording. These features show the seasonal pattern, spatial differences and missing data that the later data fusion models need to handle.

Figure 2.1 shows the VWC for the 22 sensors from September 2020 to November 2021 individually. At each location, the variation of soil moisture is an outcome of the balance between precipitation and evaporation. The VWC shows similar patterns over time for each sensor location, with high values in winter and low values in summer. The VWC vary across different sensor locations, indicating soil moisture variability throughout the Elliot Water study catchment.

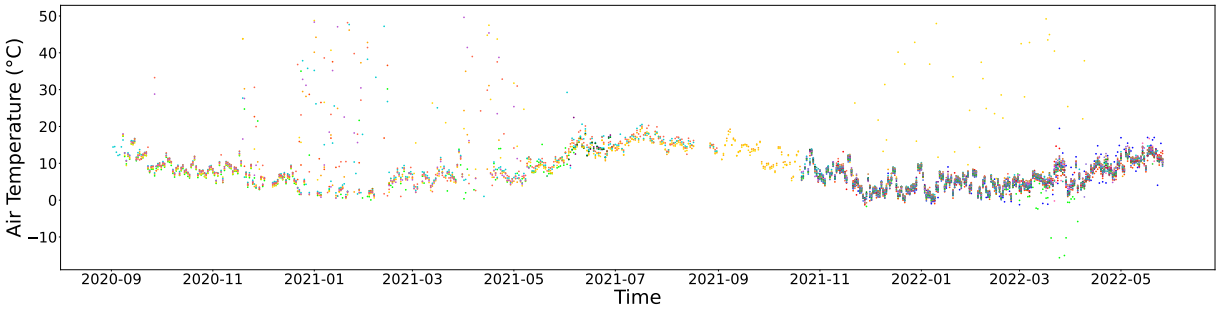


Figure 2.2: Time series plot of 15-minute air temperature for SEPA sensors in Elliot Water

Figure 2.2 shows the air temperature for the 22 sensors from September 2020 to November 2021 individually. For the air temperature, the gap between 15/08/2021 and 29/08/2021 only includes four observations, and none of them seem reasonable, so they are removed in the data preprocessing procedure. There are two large gaps during this year, so imputation may be needed for further analysis. It also shows multiple isolated unrealistically high air temperature values (up to 50 degrees Celsius). These spikes are not meteorologically plausible for the Elliot Water catchment and are most likely due to sensor errors. The spikes above 35 degrees Celsius were removed before the analysis, and since the data is recorded every 15-minute so this has a negligible impact on the overall pattern or the covariates used in the models. The plots show that the air temperature has a very similar pattern over time in each location, but does not vary much across the whole Elliot Water area. The air temperature has a strong seasonal pattern, with high values in summer and low values in winter.

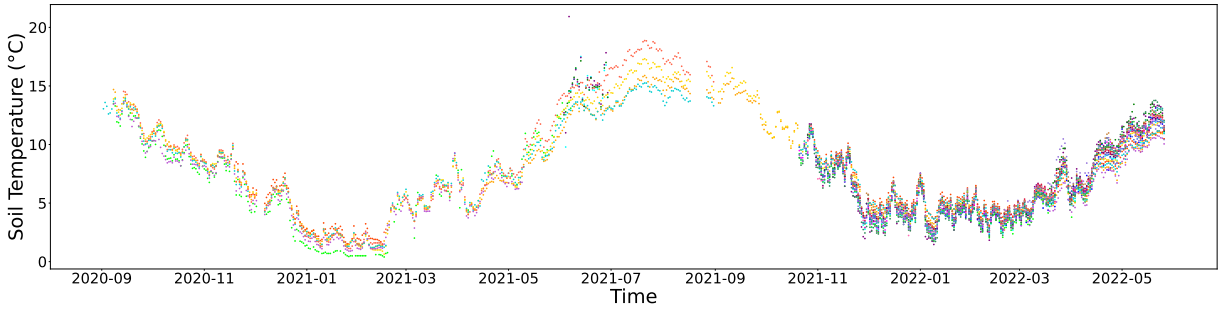


Figure 2.3: Time series plot of 15-minute soil temperature for SEPA sensors in Elliot Water

Figure 2.3 shows the soil temperature for the 22 sensors from September 2020 to November 2021 individually. A similar gap is observed in soil temperature as in air temperature, so the same processing procedure is applied. Additionally, the soil moisture data exhibits a seasonal pattern similar to that of air temperature.

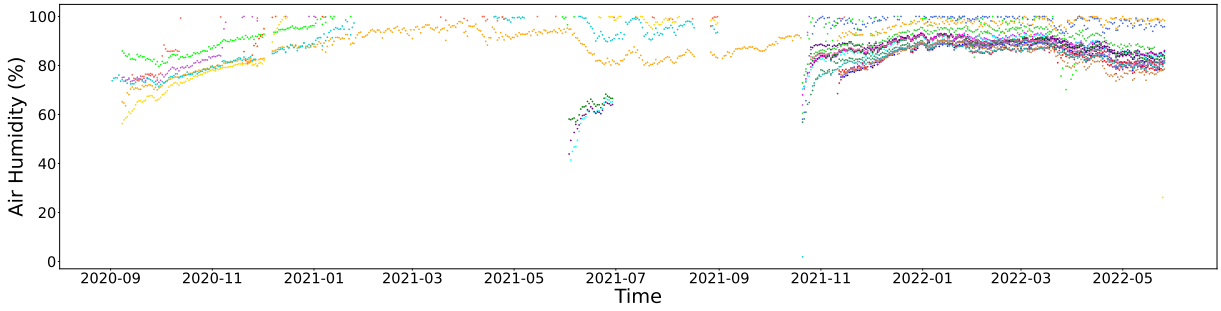


Figure 2.4: Time series plot of 15-minute Air Humidity for SEPA sensors in Elliot Water

Figure 2.4 displays the air humidity for the 7 DROPLET sensors from September 2020 to November 2021 individually. The plots do not show a very obvious pattern over time, but show a weak spatial pattern. During the pandemic, the sensors might be misbehaving because no one can get access to the field to correct them. In the humidity data, two of the sensors occasionally reset themselves, but the others remain stuck at 100%. Consequently, these saturated readings are recorded as missing values.

Figure 2.5 shows complex distribution patterns and variable relationships. The diagonal of the matrix shows the distribution of each variable (VWC, soil temperature, air temperature, and air humidity), while the off-diagonal figure shows the relationships between pairs of variables with their corresponding correlation coefficients and p-values under Kendall's tau correlation test. However, although Kendall's tau is non-parametric and more suitable for non-linear relationships, it still assumes independent observations, so the spatial-temporal autocorrelation is a concern for the spatial-temporal data. The VWC shows a left-skewed distribution with a maximum of around 40%, indicating the study catchment typically has relatively low soil moisture conditions. Soil temperature concentrates within 0-20°C, while air temperature shows a larger spread range but mostly lies in the range of 0°C to 30°C, with only a few observations above 30°C. It is noted that

air humidity readings span the entire possible range (0-100%), with a peak near 100%. The occurrence of 0% and 100% humidity readings needs careful consideration, as such extreme values might indicate potential sensor malfunction or data quality issues requiring further validation.

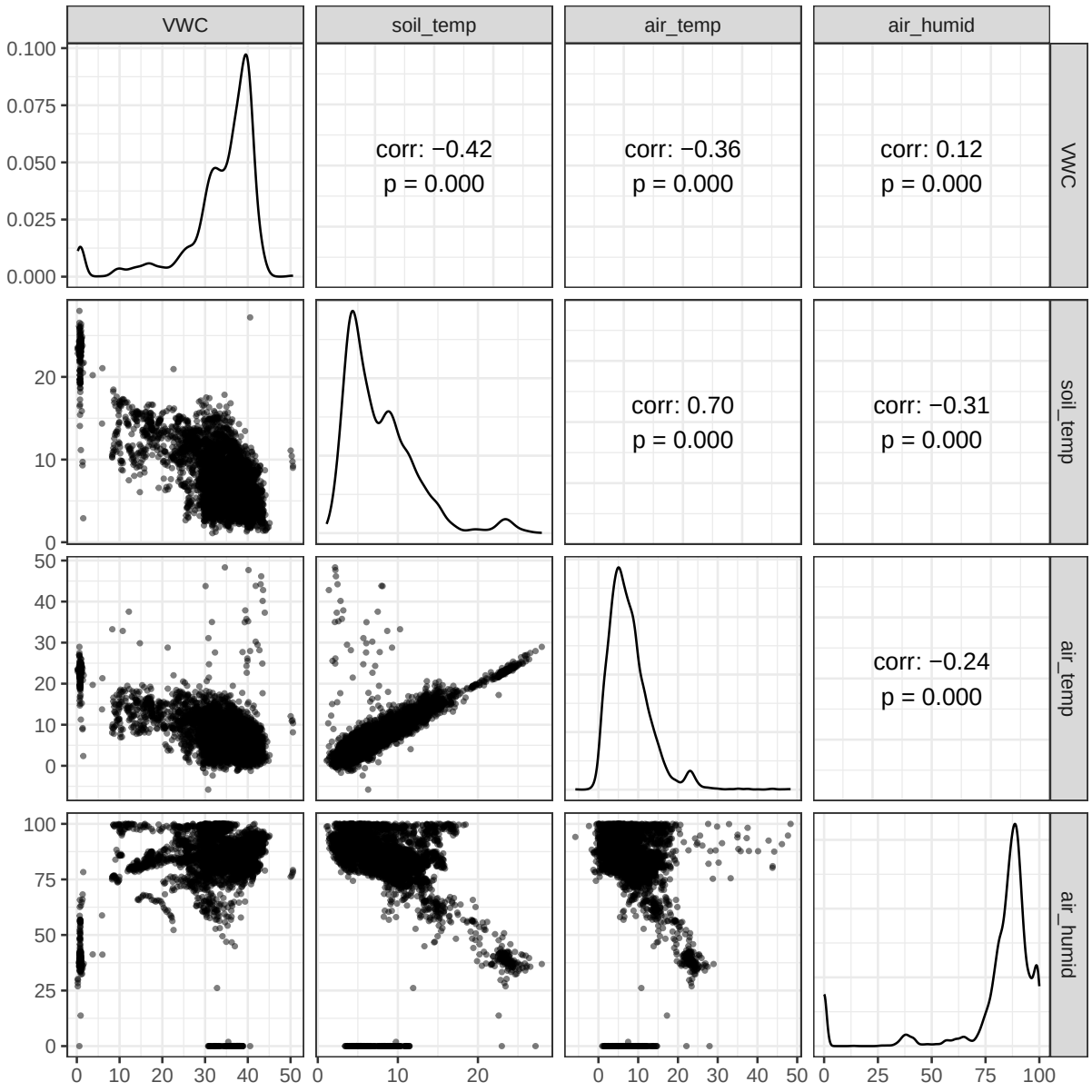


Figure 2.5: Pairwise plot for soil temperature, VWC, air temperature, and humidity.

In terms of correlations, SEPA data show different relationships. VWC shows moderate negative correlations with both soil temperature ($r = -0.42$, $p < 0.001$) and air temperature ($r = -0.36$, $p < 0.001$), which indicates consistent negative effects on soil moisture. The strongest relationship between soil temperature and air temperatures ($r = 0.70$, $p < 0.001$) shows strong consistency between ground and air conditions. Air humidity shows relatively weak correlations with other parameters; its distribution pattern and extreme values suggest the need for additional quality control measures, particularly for readings at the lower boundary of the physically possible

range. These patterns highlight the complex nature of interactions within soil moisture and other potentially related variables, where changes in one variable consistently influenced the others, but the relationships might not be linear.

2.2 COSMOS data

COSMOS sensor data, a well-maintained in-situ soil moisture dataset, is widely recognised as the benchmark in soil moisture measurements in the UK. This section introduces the COSMOS dataset by detailing its variables and examining the relationships among them. It also presents summary statistics and visualisations of the time series for each variable, with a further time series decomposition to reveal the underlying temporal patterns.

2.2.1 Summary statistics on variables

Table 2.1: Summary statistics for variables in COSMOS data: VWC, air temperature, and precipitation

Variable	Mean	SD	Min	1st Q.	Median	3rd Q.	Max	Range
VWC (%)	31.82	5.73	14.71	29.03	32.89	35.36	85.42	70.72
Air_temperature (°C)	8.74	4.63	-4.55	4.99	8.90	12.44	20.28	24.83
Precipitation (mm)	1.96	4.10	0.00	0.00	0.05	2.04	36.40	36.40

Table 2.1 shows the summary statistics for three environmental variables within the COSMOS dataset: VWC, air temperature, and precipitation. For each variable, mean, standard deviation, minimum, maximum, and quartiles are used to give an overview of the distribution of the data. Figure 2.6 shows the time series plots of VWC, air temperature and precipitation, which helps visualise the patterns in the data over time. To be specific, the missing rate of the VWC is 1%, which is unlikely to impact the overall distribution, especially if the missing is caused by some special conditions, such as sensor failure or extreme weather change. The average value of the VWC is 31.82%, with a standard deviation of 5.73%, which indicates a medium level of soil moisture with a stable fluctuation pattern. The minimum of the VWC is 14.71% while the maximum is 85.42%, which shows a large range from dry to humid soil moisture conditions and might be caused by the rainy winters and dry summers. Based on the quartiles, which are 29.03%, 32.89%, and 35.36%, most VWC data values cluster around the median, with only a few outliers. This indicates that the location has stable soil moisture levels, with moderately dry summers and wet winters.

As for the air temperature, it only has 7 missing values, so missingness has a low possibility of impacting the overall data quality. The mean temperature is 8.74°C with a standard deviation (SD) of 4.63°C, which suggests a cool climate and medium-level fluctuation. The range of the

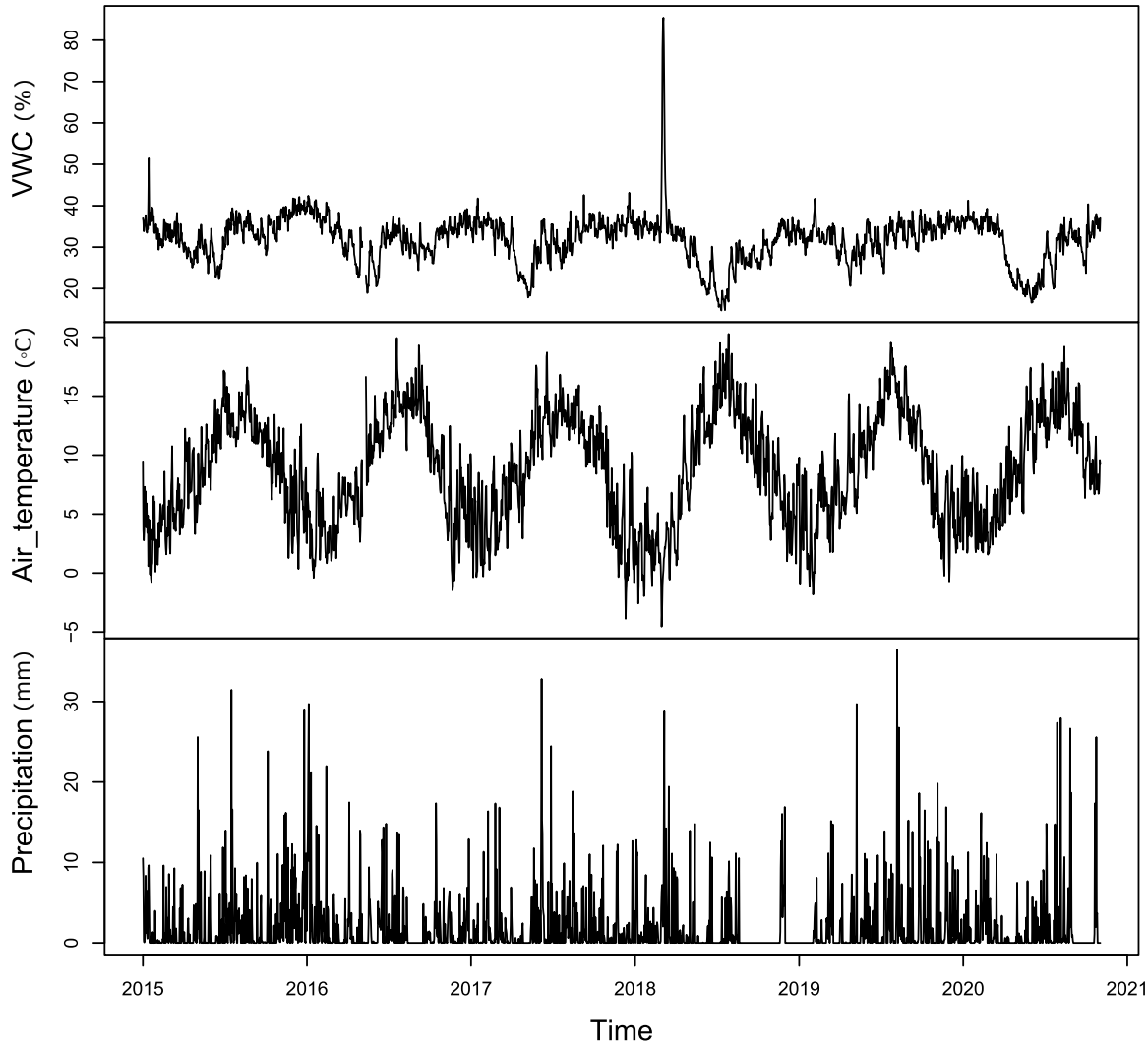


Figure 2.6: Volumetric water content (VWC), air temperature and precipitation of COSMOS data at Balruddy across 2015-2020

temperature is -4.55°C to 20.28°C, which indicates the seasonal pattern as shown in Figure 2.6, with cold winters (below zero) and warm summers. However, the air temperature doesn't include any high extremes like the VWC; it only includes some low extremes at the beginning of 2018.

There are no missing values for the precipitation. The range of the precipitation is from 0mm to 36.40 mm, which indicates that there are some days with very heavy rain, while some days have no rain at all across the year. Combined with the very low mean (1.96 mm) and high SD (4.10 mm) and quartiles (1st quantile is zero and the median is 0.05), it is confirmed that most days there is no rain or little rain, with only occasional spikes.

The distribution of the VWC seems to be slightly right skewed and air temperature seems to be normal according to the mean and standard deviation in Table 2.1 and the distribution of each variable in Figure 2.7, but the precipitation seems highly skewed according to the low median and high maximum, which suggests very infrequent rainfall events and rainfall amounts.

2.2.2 Relationship between VWC, air temperature and precipitation

Figure 2.7 shows the pairwise plots of the variables in the lower triangle, the distributions of each variable on the main diagonal, and the p-values from the Kendall Tau correlation test in the upper triangle. The missing values of VWC and air temperature are interpolated by a weighted moving average with a window size of 5 days. There is a high spike in the VWC series, and the precipitation is highly right-skewed, so a Yeo-Johnson power transformation is done to the precipitation data (Weisberg, 2001). The Yeo-Johnson power transformation is designed to handle zeros and negatives while building on the strengths of the Box-Cox power transformation. The precipitation is highly right-skewed, including zeros, so the Yeo-Johnson power transformation is used to deal with the zeros. The Yeo-Johnson power transformation is defined as:

$$\psi(\lambda, y) = \begin{cases} \left((y+1)^{\lambda} - 1\right)/\lambda & \text{if } \lambda \neq 0, y \geq 0 \\ \log(y+1) & \text{if } \lambda = 0, y \geq 0 \\ -\left[(-y+1)^{2-\lambda} - 1\right]/(2-\lambda) & \text{if } \lambda \neq 2, y < 0 \\ -\log(-y+1) & \text{if } \lambda = 2, y < 0 \end{cases}$$

Figure 2.8 shows the pairwise plot of the variables after the transformation. VWC has a moderate positive correlation with log precipitation (corr = 0.25, p = 0.000) and a moderate negative correlation with air temperature (corr = -0.25, p = 0.000). The relationship between air temperature and log precipitation is weak and statistically insignificant (corr = -0.03, p = 0.104). These findings give some insights that VWC is influenced by both temperature and precipitation, and precipitation and air temperature show little interaction.

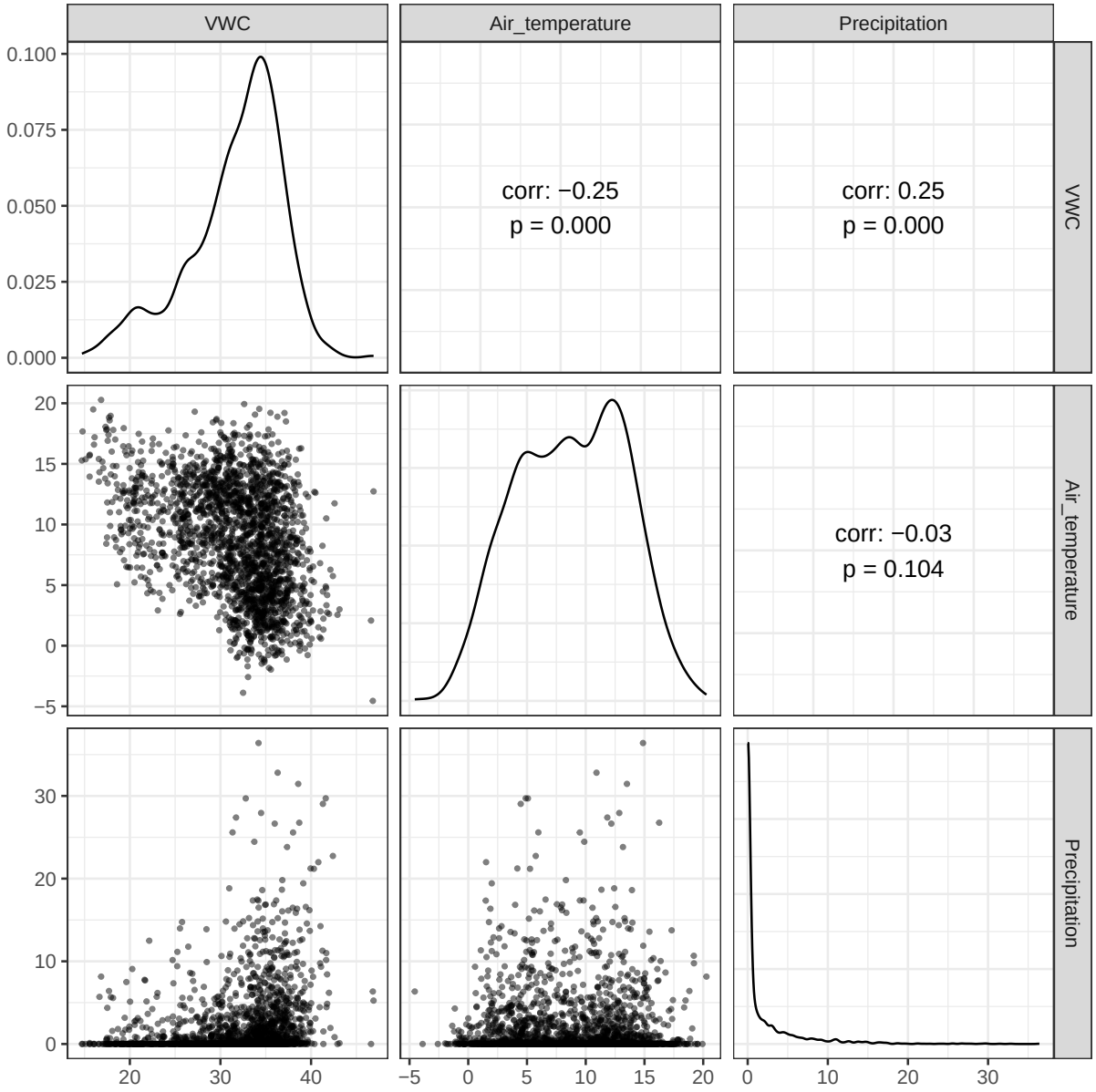


Figure 2.7: Pairwise plot showing the relationships in the COSMOS data (before transformation)

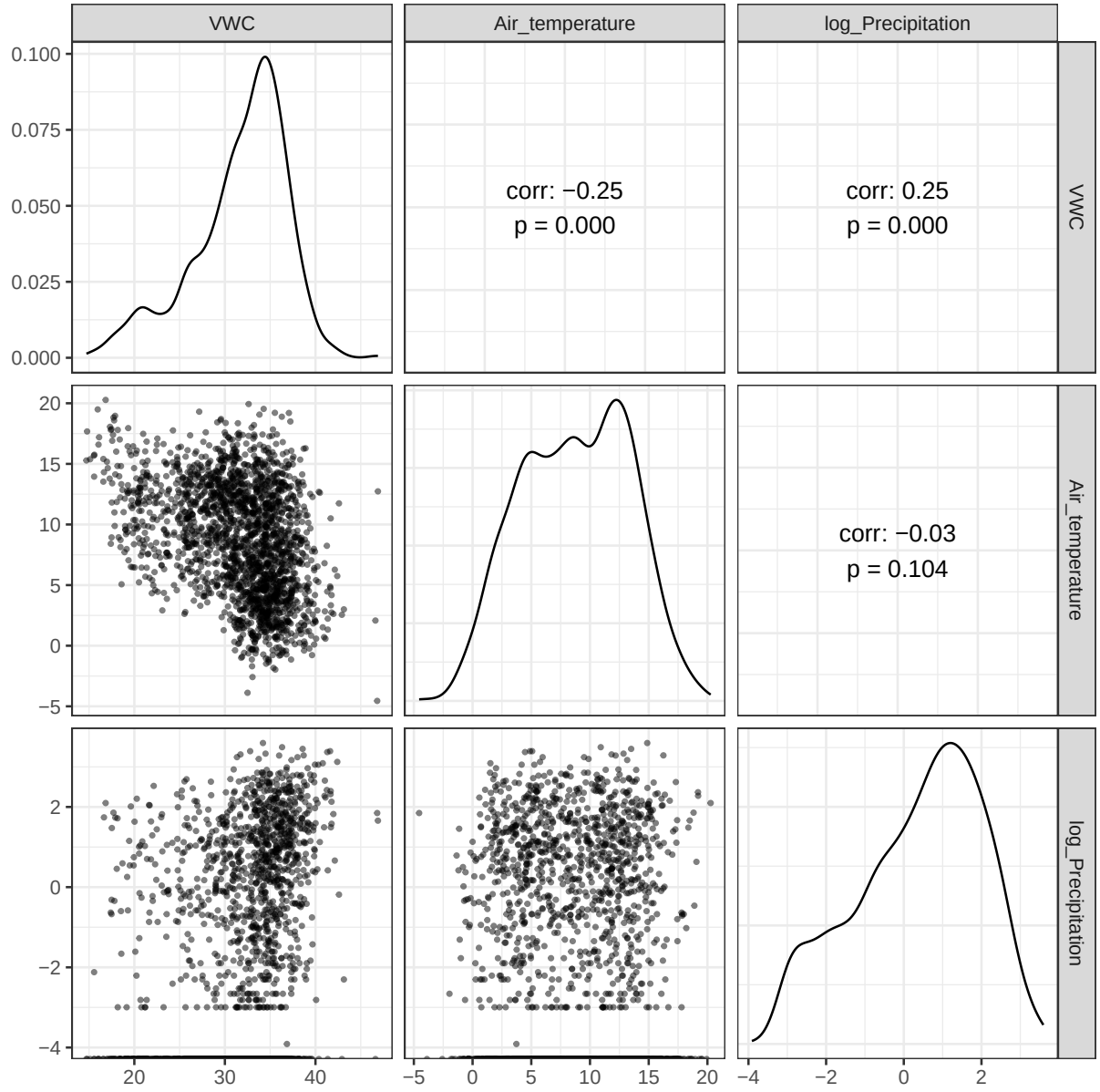


Figure 2.8: Pairwise plot showing the relationships in the COSMOS data (after transformation)

2.2.3 Time series decomposition

Time series data can show different patterns, and it is helpful to split a time series into several components, each representing an underlying pattern category. For example, a trend component, a seasonal component, and a residual component.

If we assume an additive decomposition, then the three patterns can be written as:

$$y_t = T_t + S_t + R_t,$$

where T_t denotes the trend component, which captures the long-term trend in the data; S_t

denotes the seasonal component, which represents the regular pattern in the data such as seasonal fluctuations; R_t denotes the residual component which accounts for the random variation in the data that cannot be explained by the trend or seasonal effects, such as the noise or extreme events.

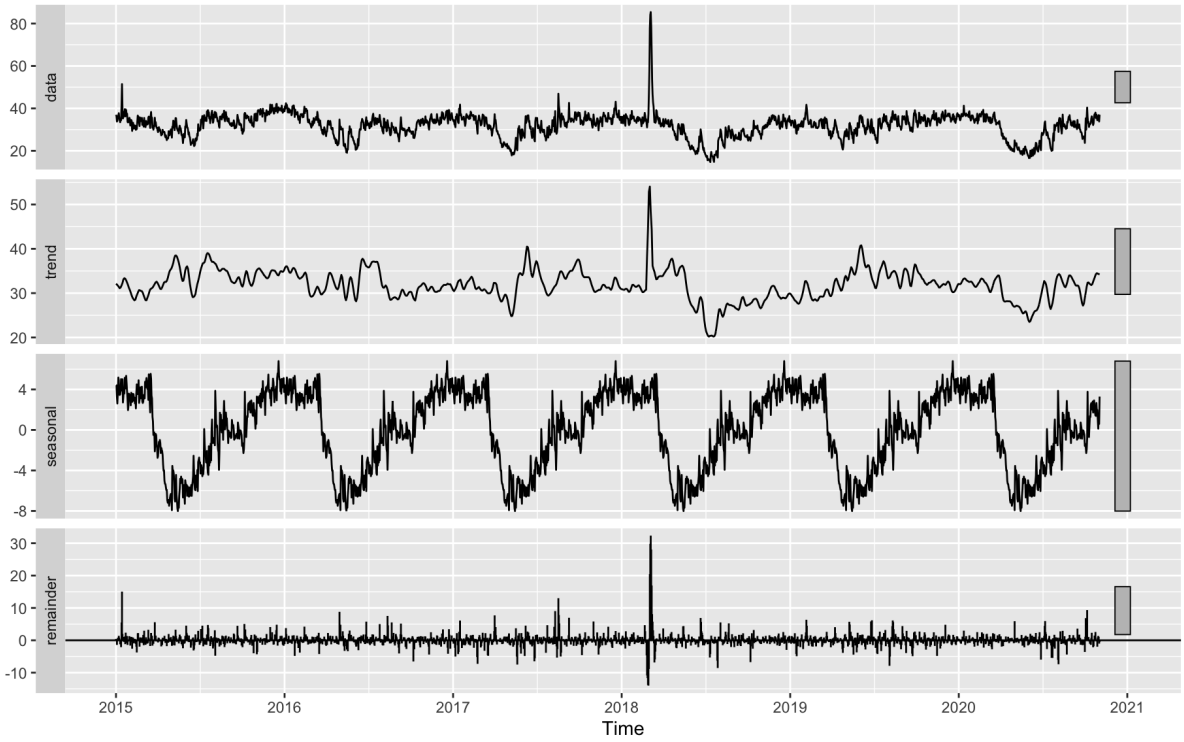


Figure 2.9: Time series decomposition of volumetric water content (VWC) from COSMOS data

Figure 2.9 and 2.10 show the time series decomposition of VWC and air temperature data from 2015 to 2020. For the VWC data, the original data show a very stable trend and fluctuation with some spikes somewhere around early 2018. The trend component shows slight fluctuations over short periods and a subtle uptrend over the five years. The seasonal component shows seasonal patterns across every year. The residual component shows mild variations with some spikes, which can not be explained by the trend and seasonal components. For the air temperature data, the 5-year VWC data show a stable trend but a mild up trend in the end. The seasonal trend is very regular, which indicates a strong seasonal pattern for each year, and the pattern is the same as the pattern in the original data. The residuals show some random noise after removing all the trends and seasonal patterns, which include many spikes over the five years. It is noted that the decomposition is only applied for the COSMOS data because they are long records with a clear annual cycle, which makes they suitable for exploring trend and seasonality. By contrast, SEPA sensors' records are shorter, so they are mainly used to explore the relationship among variables.

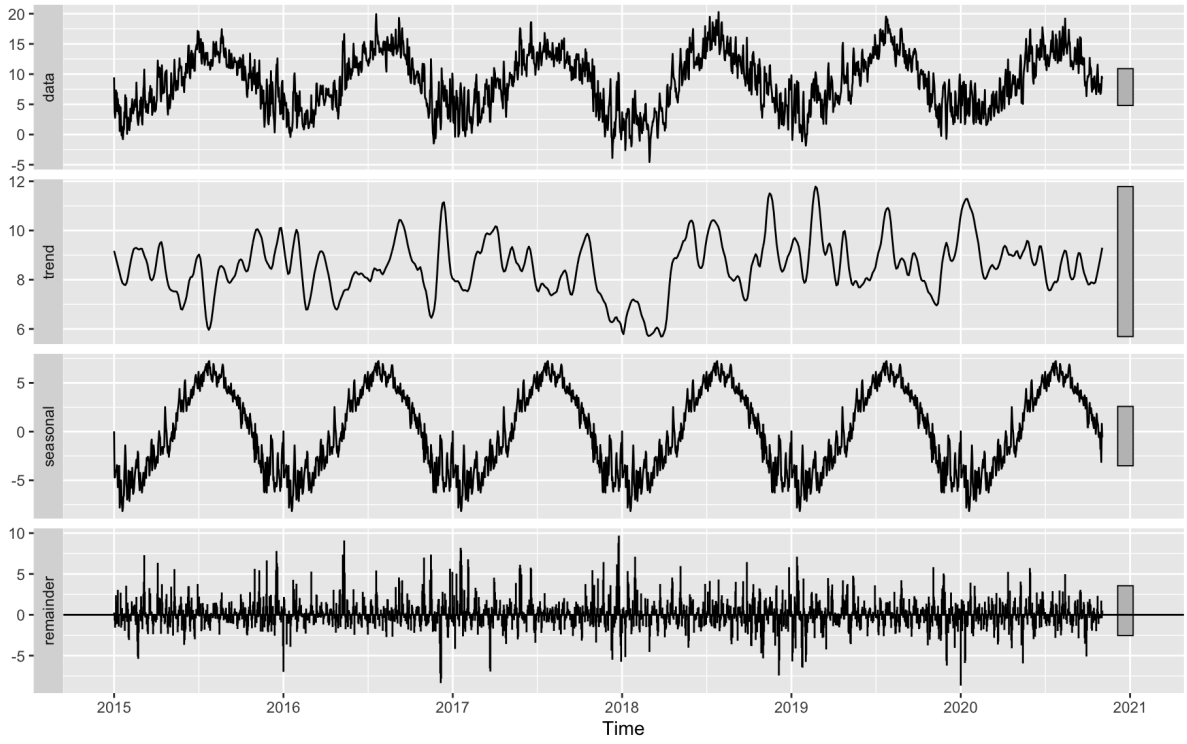


Figure 2.10: Time series decomposition of air temperature from COSMOS data

2.3 Satellite data

Copernicus satellite image data have extensive spatial coverage, making the analysis of soil moisture patterns across both spatial and temporal dimensions within the study catchment possible. Figure 2.11 demonstrates the comparison of satellite images from four different days, each representing a different season, revealing heterogeneity within the study catchment, with variability over space and time. These patterns underscore the dynamic nature of soil moisture and highlight the complexity of the underlying processes driving these patterns. To understand this variability, two modelling approaches are used here: linear models (LM) to quantify baseline relationships between variables, and generalised additive models (GAM) to capture non-linear dynamics and spatial-temporal interactions. These results provide a deep exploration of the observed patterns, bridging the gap between raw data interpretation and comprehensive understanding.

In this section, the LM and GAM are used to capture the variation in data based on longitude and latitude to investigate the spatial patterns of soil moisture. Elevation is a key factor in modelling soil moisture because it affects both local climate and hydrology. Areas at higher elevations tend to receive different amounts of rain and have different temperature patterns compared to lower areas. Elevation also influences drainage and runoff, which in turn impacts the soil's water retention. By including elevation in both LM and GAM, the model can more effectively capture these influences than by using only longitude and latitude. These models estimate the relationship between the spatial coordinates (longitude and latitude), elevation and the observed values of

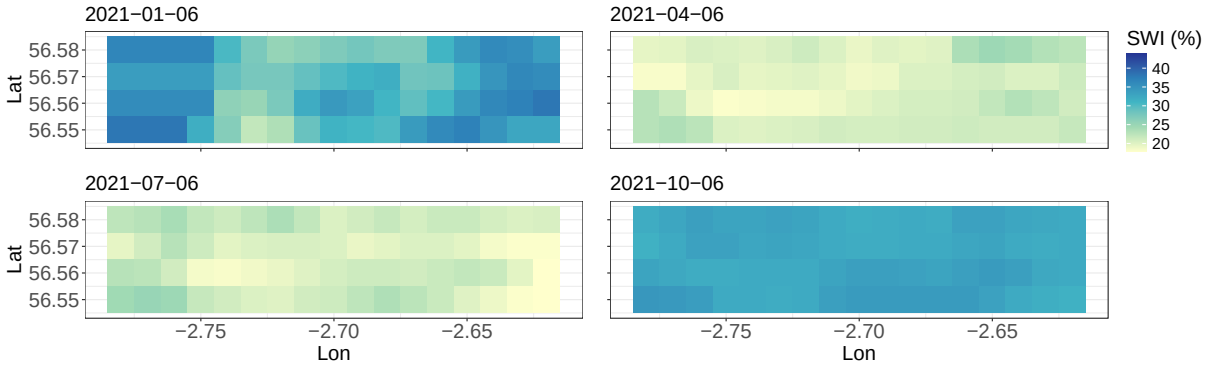


Figure 2.11: Copernicus satellite images at four selected dates 2021-01-06, 2021-04-06, 2021-07-06, and 2021-10-06. Each panel shows soil water index (SWI) values for a specific day.

SWI, allowing for the identification of geographic trends and patterns. Fitting the models to the satellite data accounts for spatial dependencies and provides insights into how the SWI changes across the Elliot water catchment. The models are fitted to data from 95 pixels on the satellite image on 06/05/2021 (Figure 1.2).

The LM is defined as follows:

$$Y_i = \beta_0 + \beta_1 \text{Lon}_i + \beta_2 \text{Lat}_i + \beta_3 \text{elevation}_i + \varepsilon_i \tag{2.1}$$

where Y_i represents the SWI value at location i, β_0 is the intercept, β_1, β_2, and β_3 are the coefficients for longitude, latitude, and elevation, respectively, and ε_i is the error term with$\varepsilon_i \sim N(0, \sigma^2)$.

Given that soil moisture patterns often exhibit complex spatial dependencies and non-linear relationships with topographic features, GAM has also been fitted to the data:

$$Y_i = \beta_0 + f_1(\text{Lon}_i, \text{Lat}_i) + f_2(\text{elevation}_i) + \varepsilon_i \tag{2.2}$$

where f_1 and f_2 are smooth functions estimated using thin plate regression splines. The bivariate function f_1 captures the spatial interaction between longitude and latitude with smoothing parameter $k = 20$, while f_2 models the non-linear relationship with elevation using $k = 10$ basis functions. The error term ε_i is assumed to be normally distributed with zero mean and constant variance.

The spatial dependency structure is accessed using variogram analysis:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(s_i + h) - Z(s_i)]^2 \tag{2.3}$$

where $\gamma(h)$ is the semivariogram value at distance h, $N(h)$ is the number of pairs of points separated by distance h, and $Z(s_i)$ represents the residual at location s_i. This analysis helps us understand the spatial correlation structure in the data and assess the models' ability to capture spatial patterns.

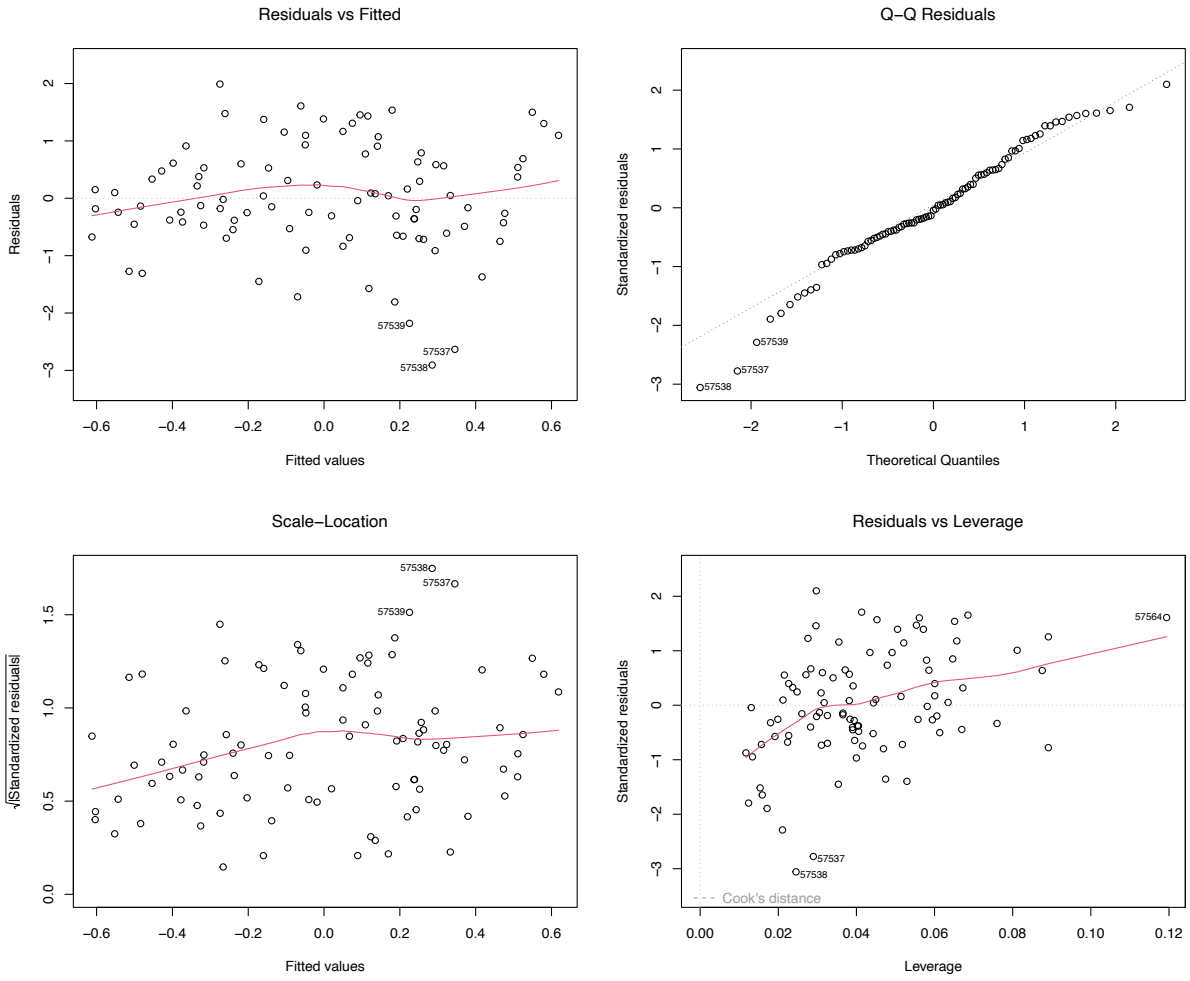


Figure 2.12: Diagnostic plots for the linear model (LM) fitted on 06/05/2021, including residuals vs. fitted values (top left) to check for non-linearity, a Q-Q plot of residuals (top right) to assess normality, a scale-location plot (bottom left) to examine the homoscedasticity of residuals, and residuals vs. leverage (bottom right) to identify influential observations and potential outliers using Cook's distance.

Figure 2.12 shows important characteristics of the LM fitted on 06/05/2021. In the Residuals vs Fitted plot (top left), there is a slight pattern with some curvature in the red smoothed line,

which suggests a potential non-linear relationship. The Q-Q plot (top right) shows slight deviations from normality, with some deviation at the tails and three outliers (points 57537, 57538, 57539). The Scale-Location plot (bottom left) shows a slight upward trend, which suggests some heteroscedasticity where variance increases with fitted values. The Residuals vs Leverage plot (bottom right) shows no observations with noticeably high leverage or Cook's distance values that would significantly influence the model fit, while points 57537 and 57538 appear as outliers. In summary, while the model satisfies assumptions well, the patterns in the residuals suggest that a more flexible modelling might be needed.

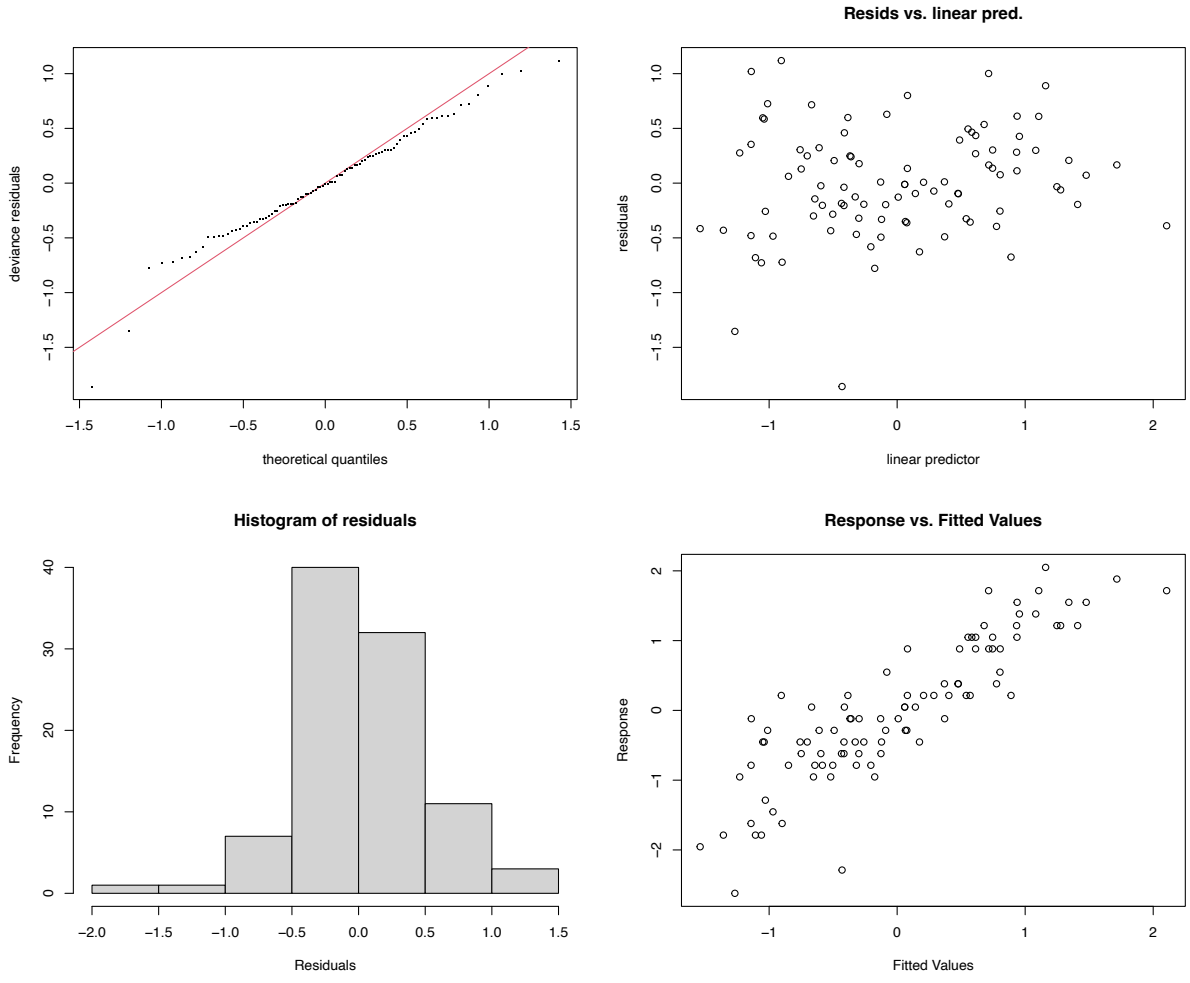
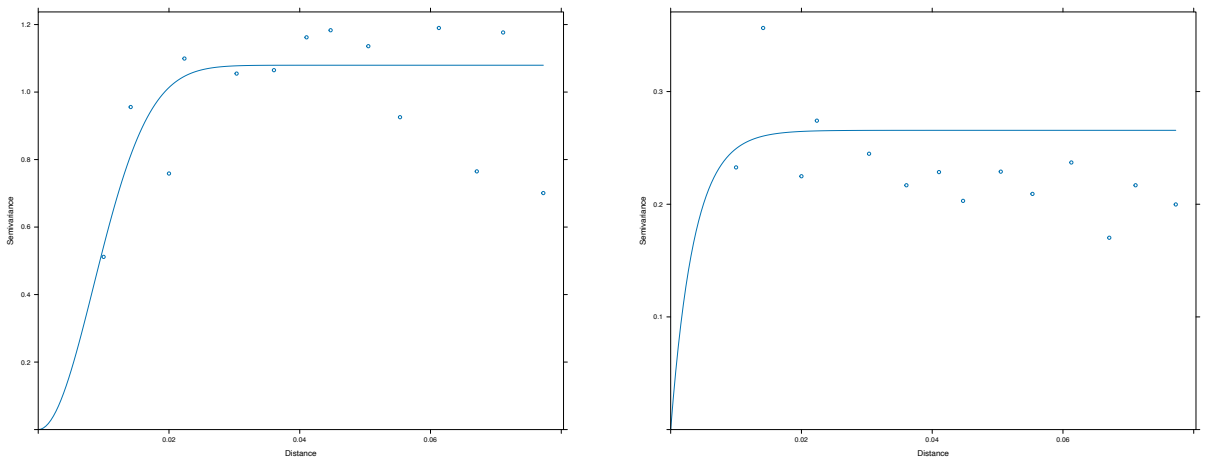


Figure 2.13: Diagnostic plots for the generalised additive model (GAM) fitted on 06/05/2021, including a Q-Q plot (top left) to assess the normality of residuals, residuals vs. linear predictors (top right) to evaluate homoscedasticity, a histogram of residuals (bottom left) to check their distribution, and a plot of response vs. fitted values (bottom right) to assess the goodness of fit and potential model bias.

The diagnostic results of GAM in Figure 2.12 show better model performance than the LM. The Q-Q plot (top left) suggests good normality of residuals with points following the red line. The residuals vs. linear predictor plot (top right) shows a random scatter pattern without

noticeable trends, which suggests that the model captures the relationships in the data reasonably well. The histogram of residuals (bottom left) appears approximately normal, though slightly right-skewed. The Response vs. Fitted Values plot (bottom right) displays a positive correlation with some scatter, indicating the model captures the main trend while showing reasonable prediction uncertainty. GAM seems to have a better performance in handling potential non-linear relationships in the satellite data.



(a) LM (b) GAM

Figure 2.14: Empirical semivariogram (points) and fitted spherical model (line) of residuals from GAM for the soil moisture data.

Figures 2.14 show the empirical semivariograms for both LM and GAM residuals, quantifying how spatial correlation changes with distance between points. For the LM (Figure 2.14a), the nugget effect is close to 0 and the sill plateaus at around 1.1, with a range of approximately 0.03 units where the semivariance levels off. The spherical variogram model fits the data very well, with empirical points following the fitted line pattern. This suggests a significant remaining spatial structure in the residuals that wasn't captured by the LM.

Figure 2.14b shows the GAM residuals variogram with different characteristics. The nugget effect remains close to 0, but the sill is lower at approximately 0.25, indicating that the GAM has captured more of the spatial variation in the data. The range is shorter as well, around 0.02 units, showing that spatial autocorrelation in the GAM residuals disappears more quickly with distance. The lower sill and shorter range in the GAM variogram suggest that this model has more effectively accounted for the spatial dependencies in the data compared to the LM.

Both variograms demonstrate that spatial autocorrelation exists in the residuals, but the GAM's lower semivariance values indicate better performance of spatial patterns. The spatial dependence extends to about 3.3 km for the LM (0.03 decimal degrees) and 2.2 km for the GAM (0.02 decimal degrees), beyond these values the observations become spatially independent.

2.4 Investigating the relationship between in-situ data and satellite data

Figure 2.15 shows the map with the sensor data and satellite data on 06/05/2021 with the EUI number for one example sensor. Figure 2.16 demonstrates the time series of selected sensors and grid from the satellite image to visualise and explore the trends, seasonality, etc in the soil moisture data. In addition, Pearson correlation, rolling correlation, and cross-correlation are employed to help understand the relationships between the two time series.

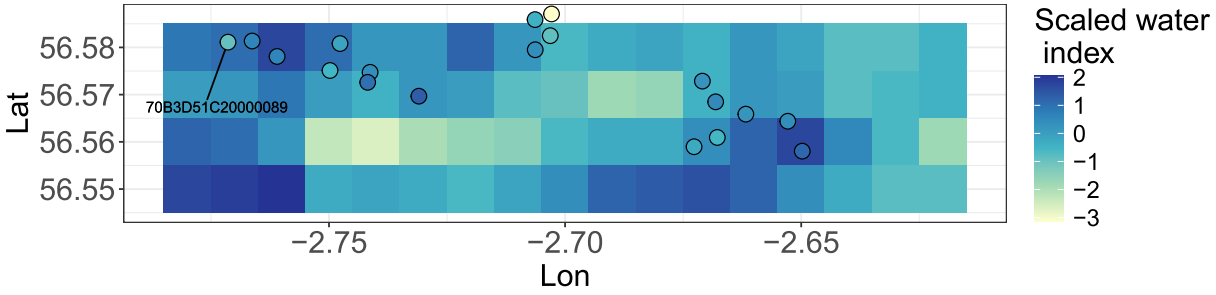


Figure 2.15: A map with the sensor data and satellite data on 06/05/2021.

The 15-minute VWC is averaged to daily data to have the same temporal resolution as SWI. Figure 2.16 shows the time series of scaled SWI in orange and VWC in blue. Both variables fluctuate throughout time, with SWI showing more variability than VWC. There is a period from April to July 2021 where both SWI and VWC drop significantly, indicating a potential dry period. At the end of 2021, both variables increased, which is possibly due to seasonal changes or large precipitation. The VWC and SWI data show similar time series trends across all locations in the long term, but not in the short term. This suggests that modelling the short-term correlation may require further consideration.

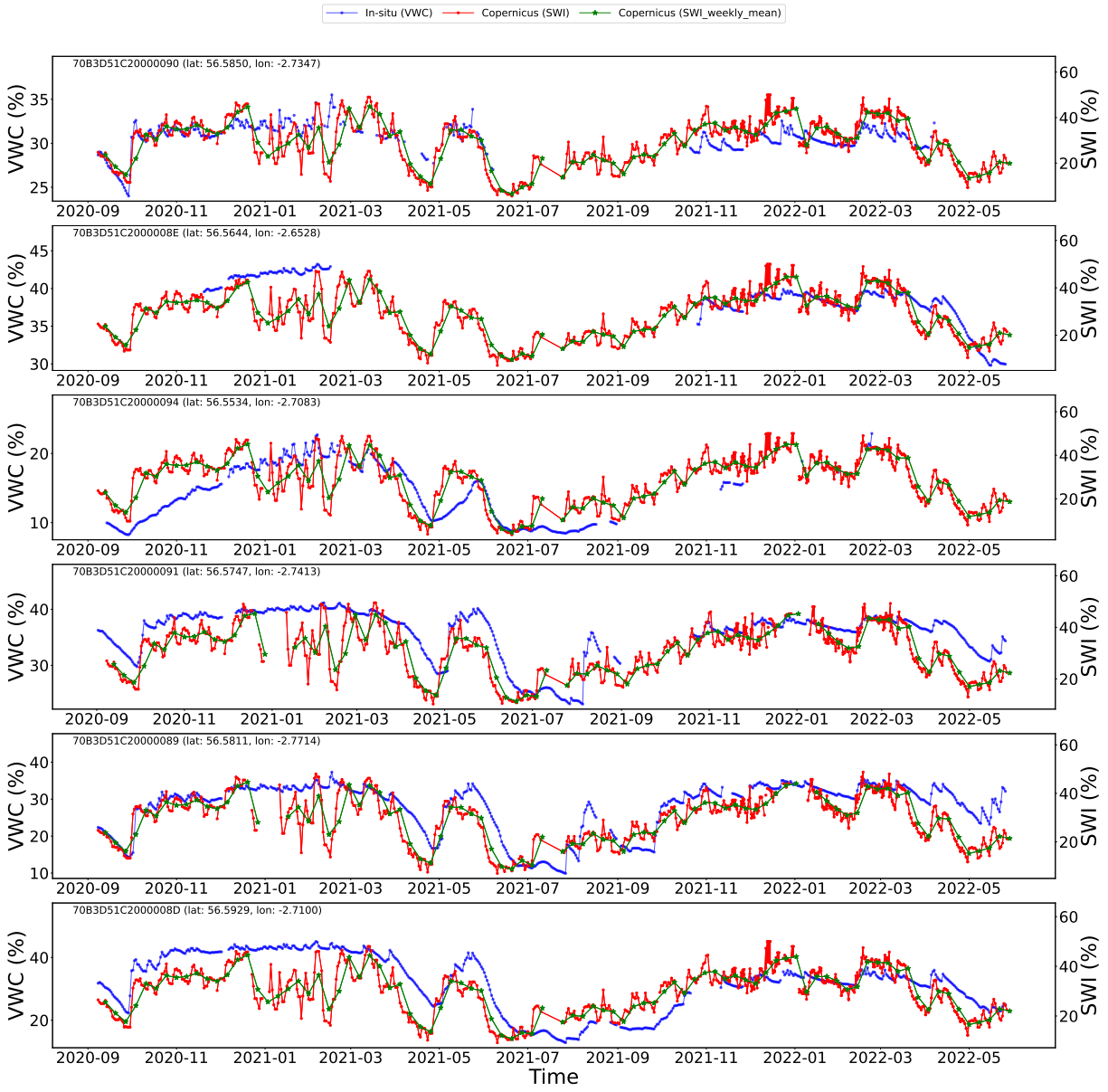


Figure 2.16: Time series of in-situ volumetric water content (VWC; left *y*-axis) and Copernicus Soil Water Index (SWI; right *y*-axis) for six example SEPA sensors in the Elliot Water catchment, from September 2020 to May 2022. In each panel, the blue line shows daily mean VWC (%), obtained by aggregating 15-minute sensor readings and expressing VWC on a 0–100% scale. The red line shows the collocated Copernicus SWI (%), extracted at the 1 km grid cell covering each sensor location on its native 0–100 index scale, and the green line shows a 7-day moving average of SWI to highlight slower temporal variation. Panels are labelled by sensor ID and geographic coordinates (latitude, longitude).

2.4.1 Check stationarity

Non-stationary time series (with trends/seasonality) will create misleading correlations. So the Augmented Dickey-Fuller (ADF) test is used to test the stationarity of the time series, and if it is not stationary, differencing or transformations may need to be applied to the time series before

any further analysis (Details of the ADF are provided in Section 1.4.4).

The null hypothesis of the ADF test is that the time series has a unit root (i.e., non-stationary). A more negative ADF statistic suggests stronger evidence against the presence of a unit root, implying that the series is stationary. A low p-value (<0.05) implies stationarity. The ADF test is typically performed with different lag lengths to account for serial correlation. The critical values for significance are compared with the computed ADF statistic to determine stationarity.

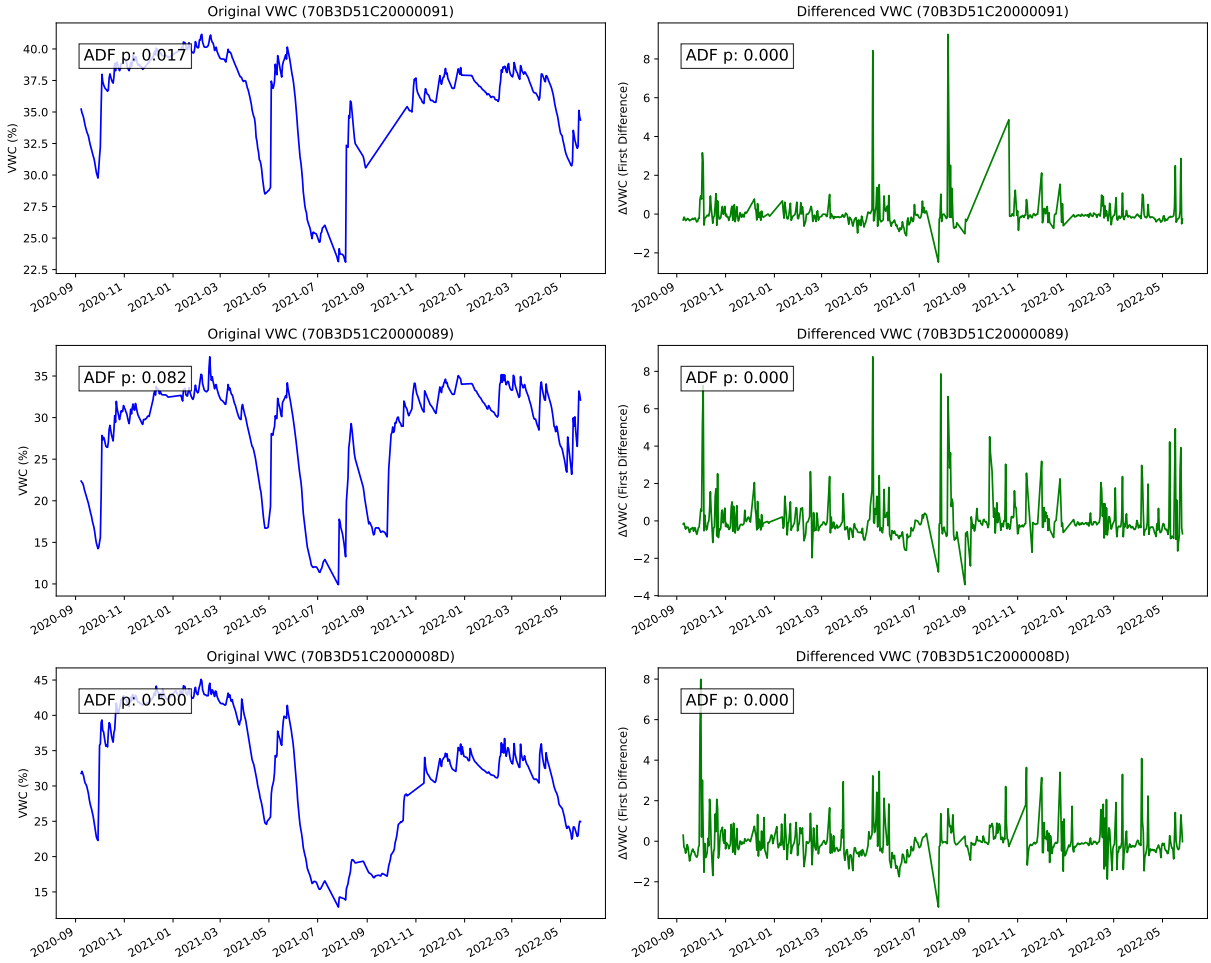


Figure 2.17: The left column displays the SEPA sensor data: original volumetric water content (VWC) time series for three locations, while the right column shows the first-order differenced series.

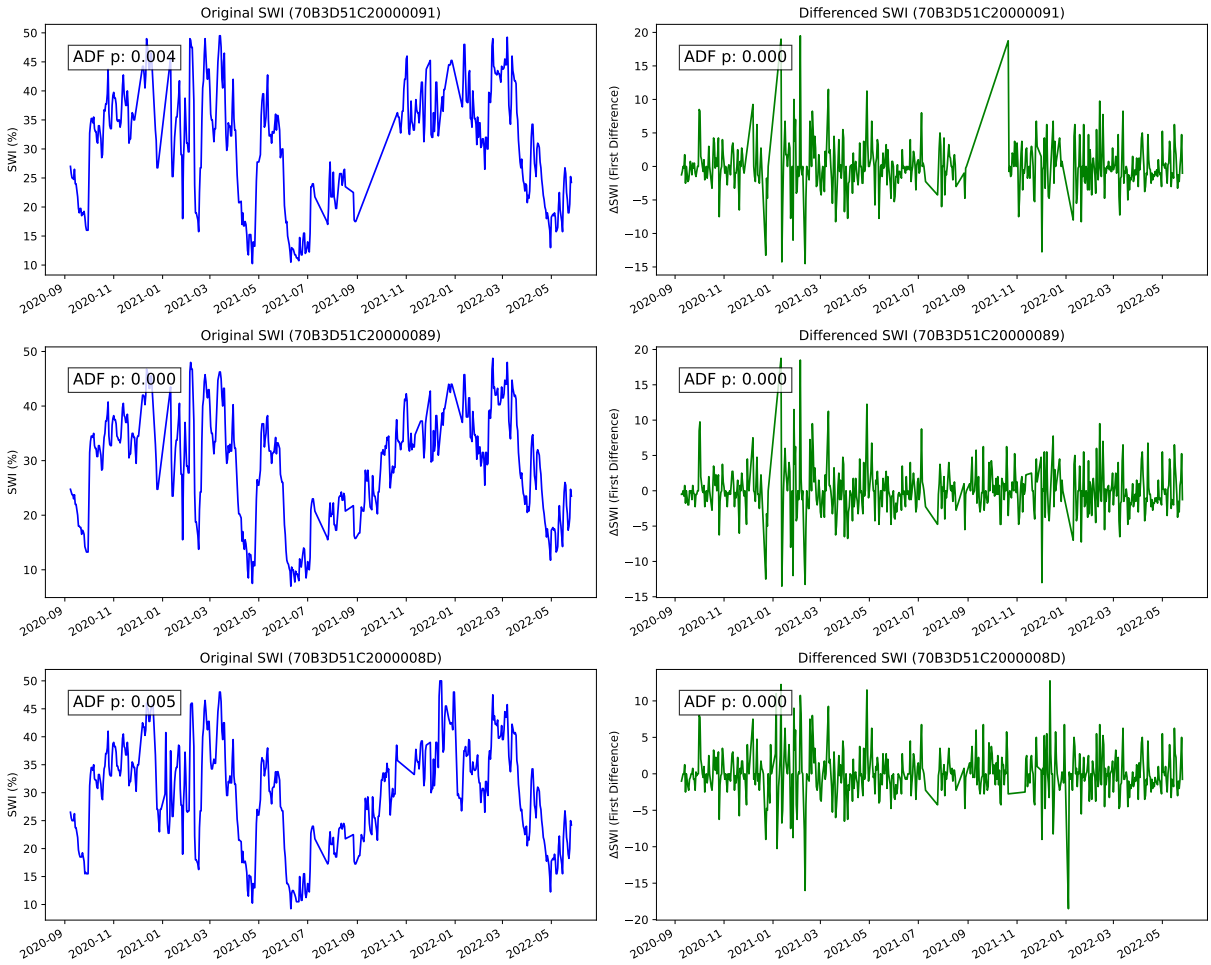


Figure 2.18: The left column displays the satellite data: original soil water index (SWI) time series for three locations, while the right column shows the first-order differenced series.

Figure 2.17 and Figure 2.18 present the VWC and SWI time series stationarity analysis for three different locations (focusing on three locations with minimal gaps to ensure data quality), using the ADF test and differencing. Both time series show clear trends and fluctuations over time, indicating non-stationarity in the time series. The ADF test p-values for the original SWI are 0.004, 0.000, and 0.005, respectively. Since these p-values are below the significance level (0.05), the null hypothesis of the ADF test is rejected, suggesting that the original SWI series is weakly stationary. However, based on the visualisation, some non-stationarity might still exist, and differencing is applied for further analysis. The differenced series seems stationary, with fluctuations centred around zero and fewer trends. The ADF test p-values drop to 0.000 for all differenced series, strongly confirming stationarity. The original VWC time series for each location shows trends and fluctuations, suggesting potential non-stationarity. The ADF test for the p-values of the original VWC series are 0.017, 0.082 and 0.500, respectively. The top and the middle ones with p-values of 0.017 and 0.082 have p-values slightly above 0.05, which means they might be weakly stationary but still have some trends in visualisation. The bottom one with 0.500 p-values indicates strong non-stationarity. The differencing is applied to all the original VWC time series to remove the trends and make the time series stationary for further

study. The ADF test p-values drop to 0.000 for all differenced VWC series, which confirms stationarity. The Augmented Dickey-Fuller (ADF) test results indicate that the original series has some non-stationary characteristics, which are removed after differencing, making the series suitable for further modelling.

2.4.2 Temporal Autocorrelation in VWC and SWI

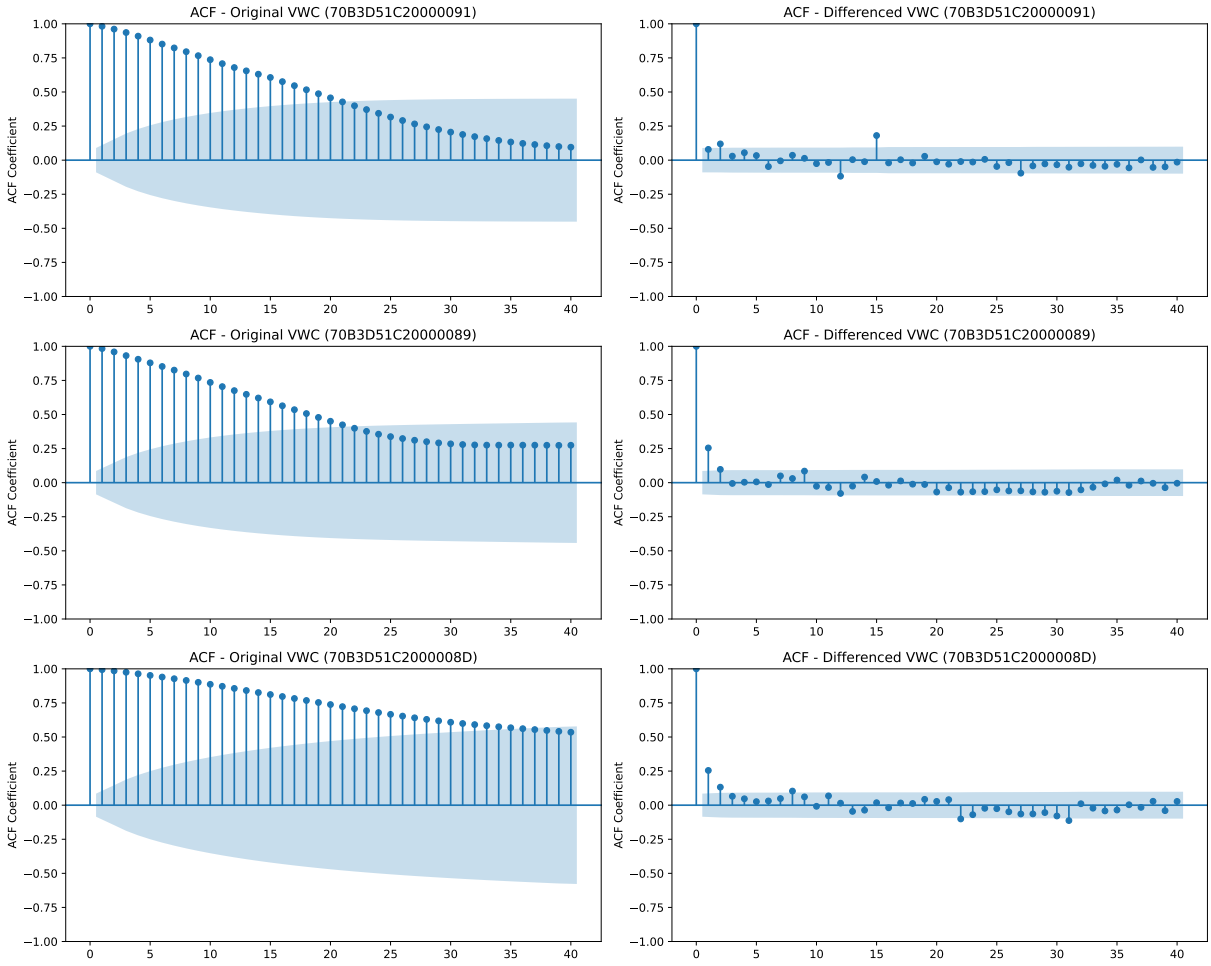


Figure 2.19: ACF for original (non-differenced) VWC and differenced VWC

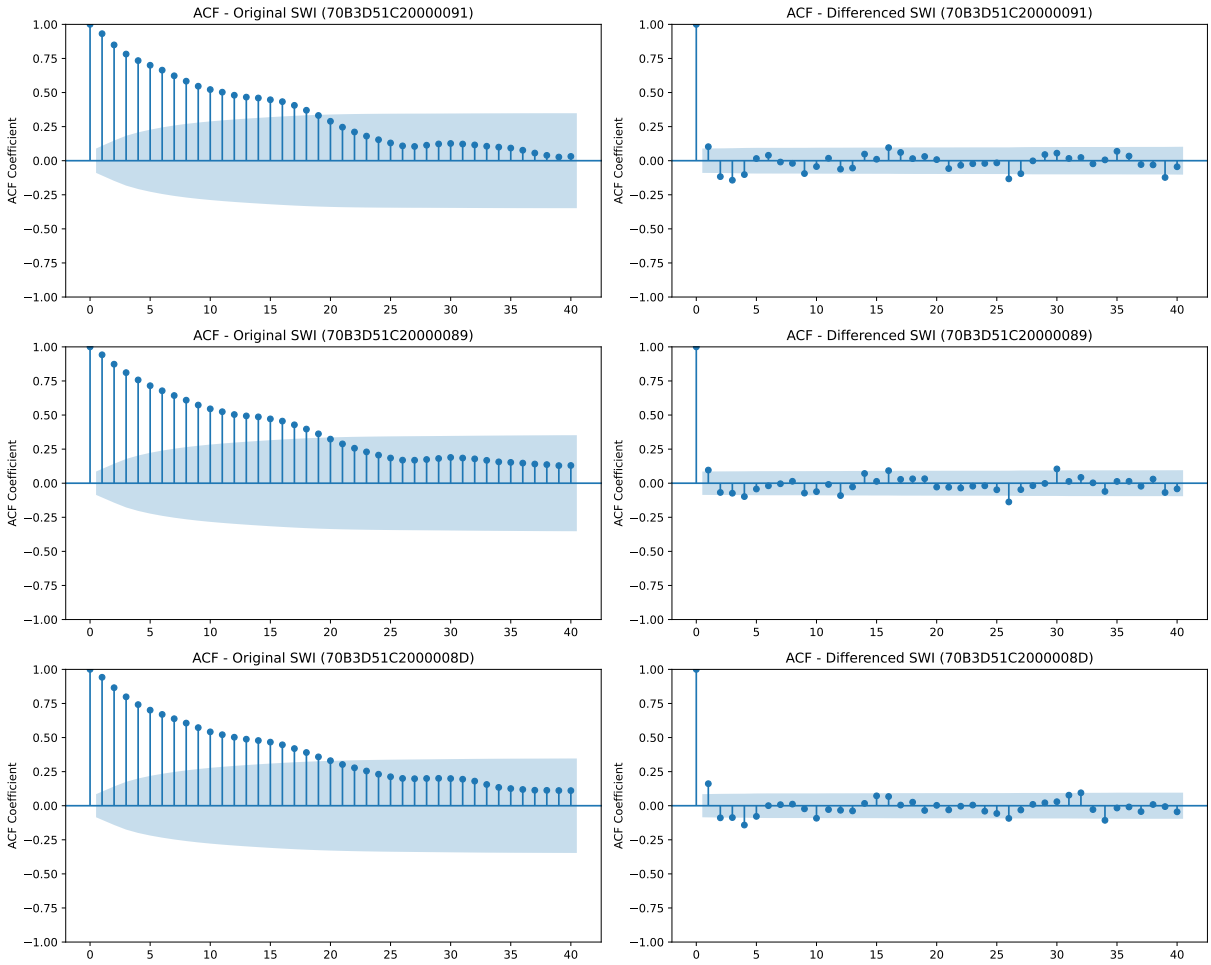


Figure 2.20: ACF for original (non-differenced) SWI and differenced SWI

Figure 2.19 and 2.20 reveal different temporal autocorrelation patterns in VWC and SWI. For the original (non-differenced) VWC, the autocorrelation coefficients decay slowly over lags, indicating long-term temporal autocorrelation. This aligns with the high ADF p-values (>0.05), which confirms that the original VWC data have trends. In contrast, differenced VWC (ΔVWC) shows a sharp drop in autocorrelation after lag 0, with coefficients of most lags remaining within the credible intervals. This suggests that first-differencing removes the trend, resulting in a short-term autocorrelation stationary series. Similarly, the original SWI exhibits autocorrelation over many lags, but decays faster, which indicates a short memory and a greater reaction to the environmental factors. It is noted that the first-differencing is used here only for exploratory purposes, and it is not used in subsequent modelling.

2.4.3 Pearson correlation

When exploring the relationship between two time series, the Pearson correlation is often used because it helps quantify how closely the two series move together over time. It assumes that the relationship between the two series is linear, which means that if one series increases, the other one will increase (or decrease). Pearson correlation can indicate whether the series has a

consistent relationship over time and the strength of the relationship between the two series. The Pearson correlation coefficient is defined as follows:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} \cdot \sqrt{\sum(Y_i - \bar{Y})^2}}, \tag{2.4}$$

where: X_i, Y_i are individual data points of the variables X and Y, $\bar{X}, \bar{Y}$ are the mean values of X and Y, respectively. The numerator represents the covariance between X and Y. The denominator normalises the expression using the standard deviations of X and Y.

The Pearson correlations between VWC and SWI are 0.61, 0.70, and 0.72 at the three sites, indicating varying relationships between the two variables across different locations. The highest correlation (0.72) is for sensor 70B3D51C20000089, suggesting a strong relationship between VWC and SWI at this location. In contrast, the lowest correlation (0.61) is at 70B3D51C2000008D, indicating a weaker but still positive linear relationship. These differences reflect local soil moisture patterns, sensor accuracy, or environmental factors influencing soil moisture dynamics. While Pearson correlation provides insight into the linear relationship, it does not capture potential nonlinear dependencies. To gain a more comprehensive understanding of the relationship between VWC and SWI, additional methods such as rolling-window correlation and cross-correlation analysis are needed. Further investigating the dynamics between VWC and SWI needs additional methods. For example, rolling-window correlation can be applied to assess how the linear relationship evolves over time, and cross-correlation analysis can investigate the lag effect of the two time series.

2.4.4 Rolling correlation

Rolling correlation is a localised measure computed over a moving window, capturing short-term relationships between the two time series rather than relying on a single global estimate. So even if the overall time series is non-stationary, a rolling correlation can still show how their relationship evolves over time. However, if both time series have strong trends, they may be highly correlated even if there is no real underlying relationship. Differencing can help address this issue. The rolling correlation over moving windows (15 days) will be computed using the differenced time series to ensure more reliable results.

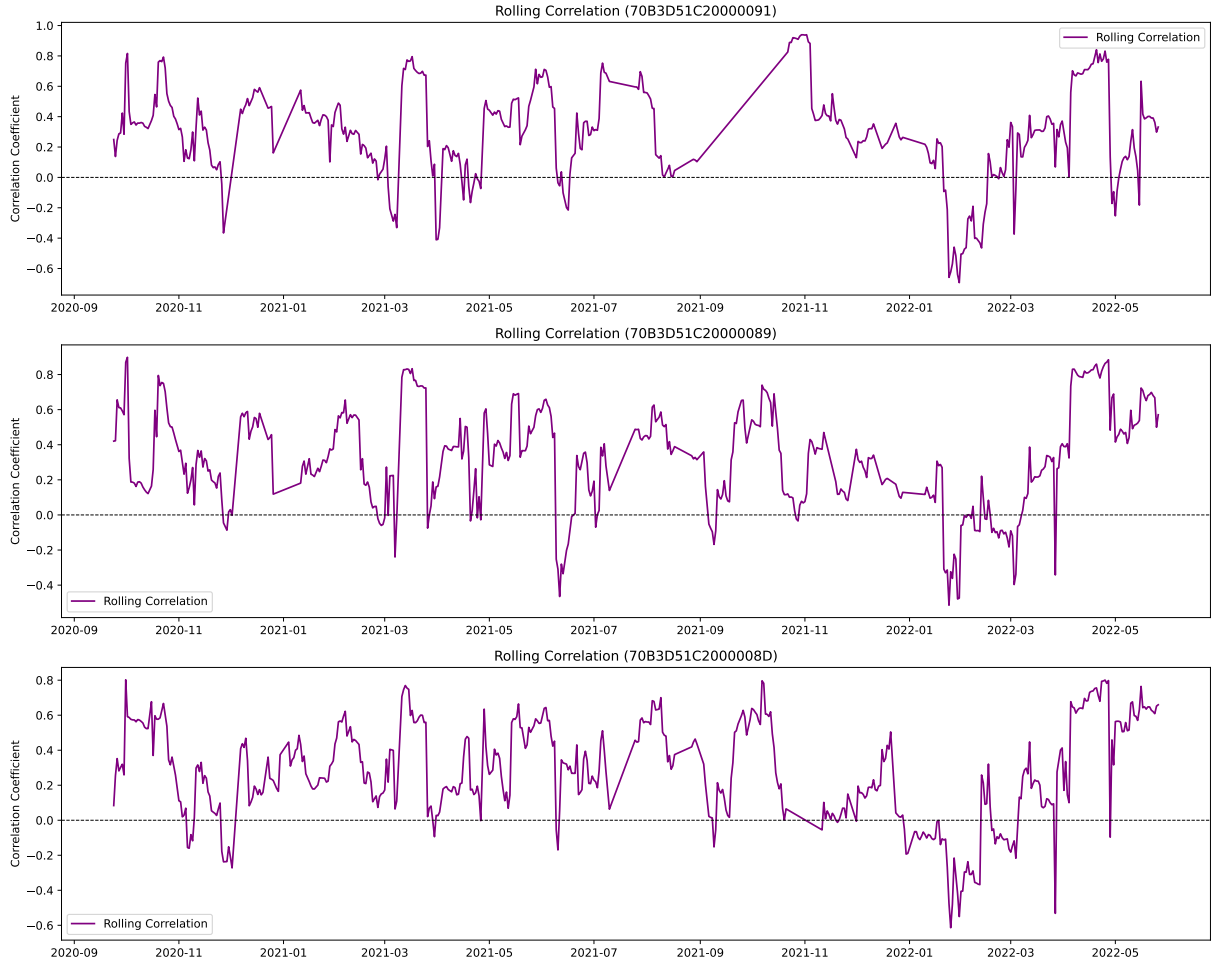


Figure 2.21: Rolling cross correlation over moving windows (15 days) between volumetric water content (VWC) and soil water index (SWI).

Figure 2.21 illustrates how the relationship between VWC and SWI evolves over time for different locations. Since both VWC and SWI are soil moisture measurements at the same location, the rolling correlation results provide insights into how well these two time series track each other over time. The rolling correlation fluctuates over time rather than staying constant, suggesting that VWC and SWI do not always behave in the same way. This could be due to differences in how each equipment responds to environmental factors such as rainfall or soil type. When the correlation is high, it indicates that VWC and SWI are closely aligned, meaning both data streams are capturing similar soil moisture dynamics. A decrease in correlation suggests that one measurement is reacting differently from the other one, probably due to sensor differences or differences in data processing for in situ sensors and Copernicus satellite data. Negative correlations suggest that one variable is increasing while the other is decreasing, which could happen if SWI, which is derived from the radar sensor, lags behind real soil conditions measured by VWC as a ground-based measurement. VWC is a direct In-situ measurement, whereas SWI is derived from remote sensing, which might introduce noise. If one sensor has a different response time (e.g., SWI integrating moisture over a larger area or with a lag), this could explain some of the fluctuations. Seasonal changes might cause shifts in correlation. For example, in dry

conditions, VWC might respond more quickly to evaporation, while SWI has variations. Sudden decreases could be caused by localised factors like sensor errors or differences in soil types.

2.4.5 Cross-correlation (lagged analysis)

Serial correlation affects the correlations among variables measured in time, so it is important to recognise the serial correlation before further analysis. Cross-correlation analysis is used to investigate the lag effects between VWC and SWI, if there are any.

2.4.6 Serial correlation

The cross-covariance coefficients between x_t and y_t series at lag+k is defined as:

$$\gamma_{xy}(k) = E\left[(x_t - \mu_x)\left(y_{t+k} - \mu_y\right)\right] \quad k = 0, 1, 2, \ldots \tag{2.5}$$

where $\mu_x = E[x_t]$ denotes mean of x_t, $\mu_y = E[y_t]$ denotes mean of y_t, and y_{t+k} denotes the value of y_t shifted by k lags. $k > 0$ measures how x_t relates to future values of y_t (e.g., y_{t+1}, y_{t+2}), $k < 0$ measures how x_t relates to past values of y_t (e.g., y_{t-1}, y_{t-2}), $k = 0$ is equivalent to the standard covariance between x_t and y_t.
Cross-covariance is often normalised to compute the cross-correlation coefficient:

$$\rho_{xy}(k) = \frac{\gamma_{xy}(k)}{\sigma_x \sigma_y}, \tag{2.6}$$

where σ_x and σ_y are the standard deviations of x_t and y_t, respectively.

The cross-correlation coefficients of VWC and SWI are computed as follows:

$$r(k) = \frac{\sum(VWC_t - \bar{VWC})(SWI_{t+k} - \bar{SWI})}{\sqrt{\sum(VWC_t - \bar{VWC})^2 \sum(SWI_{t+k} - \bar{SWI})^2}}, \tag{2.7}$$

where VWC_t and SWI_t are the time series values at time t, $\bar{VWC}$ and $\bar{SWI}$ are the mean values of each time series and k represents the lag.

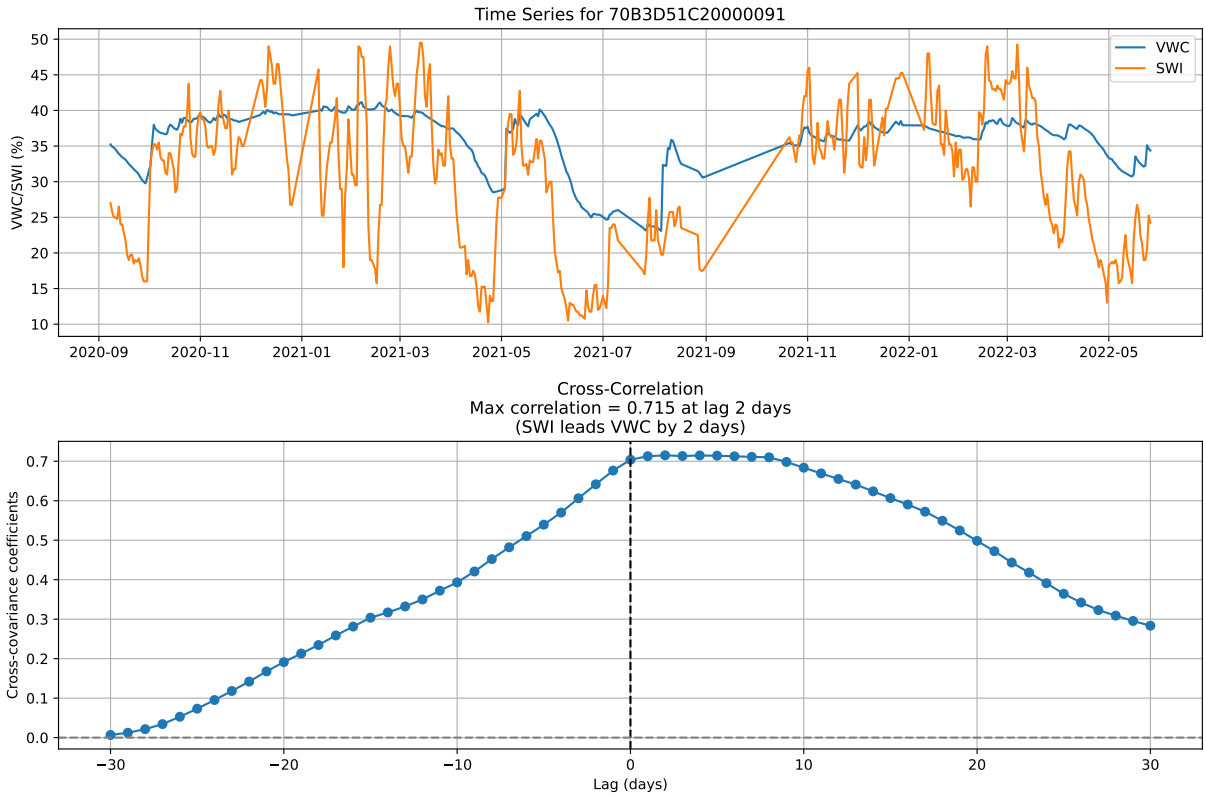


Figure 2.22: Cross-correlation between Volumetric Water Content (VWC) and Soil Water Index (SWI). The top panel shows the time series of VWC and SWI measurements. The bottom panel presents the cross-correlation analysis between the stationary time series at different time lags, where positive lags indicate SWI leading VWC and negative lags indicate VWC leading SWI. The maximum correlation coefficient (r = 0.715) occurs at lag 4 days, suggesting that SWI leads VWC by 2 days.

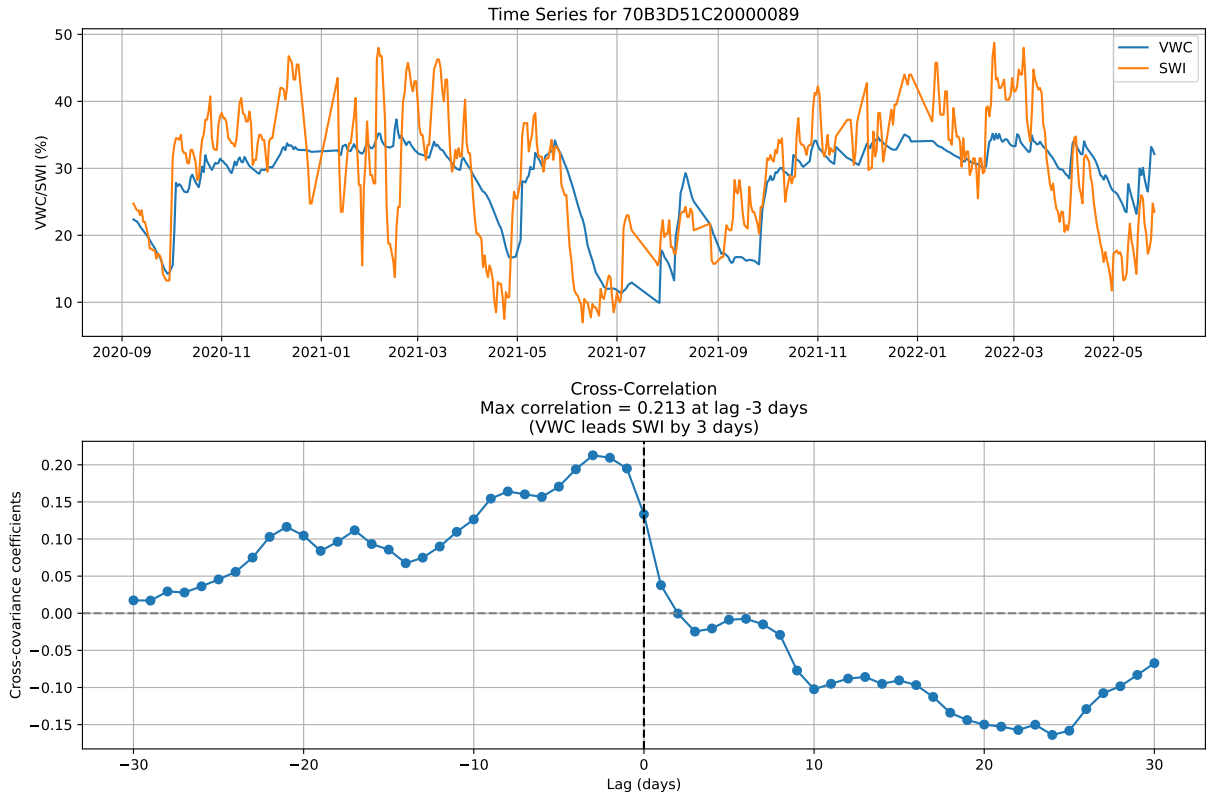


Figure 2.23: Cross-correlation between Volumetric Water Content (VWC) and Soil Water Index (SWI). The top panel shows the time series of VWC and SWI measurements. The bottom panel presents the cross-correlation analysis between the stationary time series at different time lags, where positive lags indicate SWI leading VWC and negative lags indicate VWC leading SWI. The maximum correlation coefficient (r = 0.213) occurs at lag -3 days, suggesting that VWC leads SWI by 3 days.

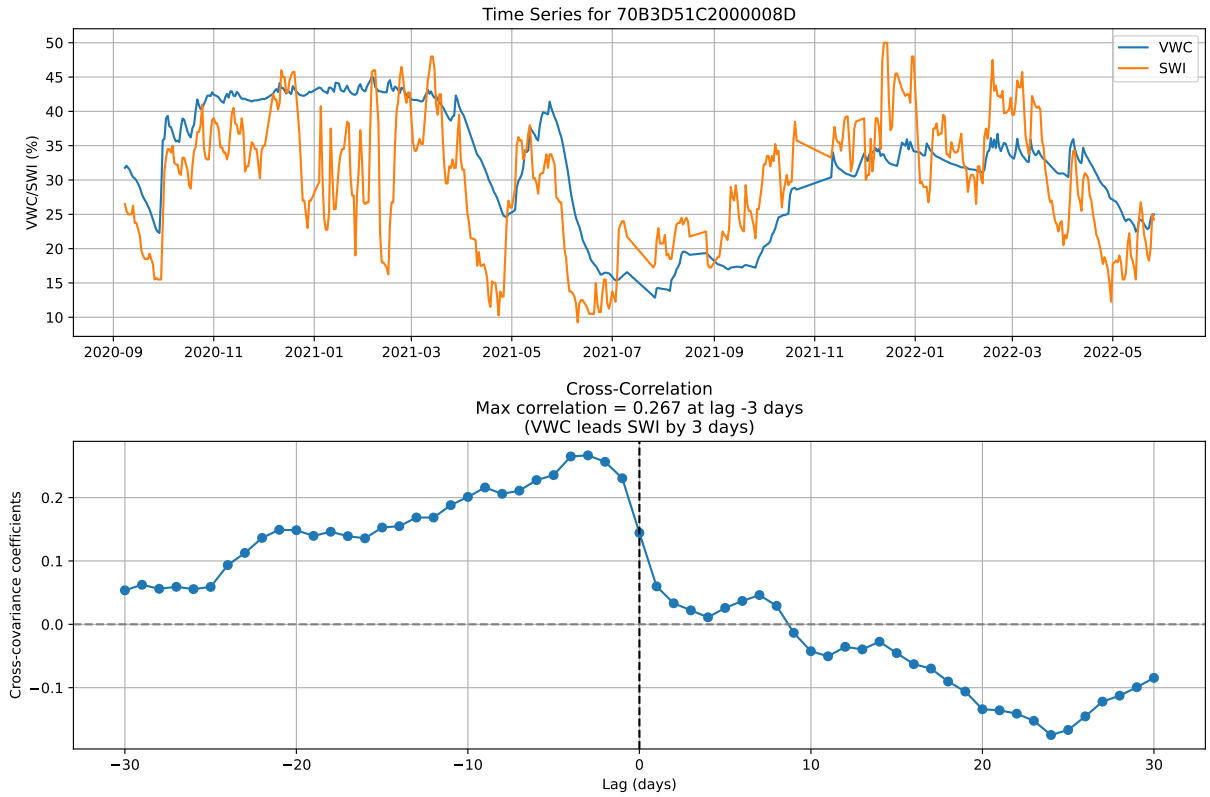


Figure 2.24: Cross-correlation between Volumetric Water Content (VWC) and Soil Water Index (SWI). The top panel shows the time series of VWC and SWI measurements. The bottom panel presents the cross-correlation analysis between the stationary time series at different time lags, where positive lags indicate SWI leading VWC and negative lags indicate VWC leading SWI. The maximum correlation coefficient (r = 0.267) occurs at lag -3 days, suggesting that VWC leads SWI by 3 days.

Figure 2.22, 2.23, 2.24 show the correlation between the VWC and SWI in different locations at different time lags. For example, in Figure 2.24, the correlation is generally positive for negative lags, which means that in this specific location, the changes in the VWC tend to occur before the SWI. The correlation increases as the lag approaches the maximum lag. At lag 0, VWC and SWI are aligned in time, and they drop sharply after lag 0, becoming negative for larger positive lag values. This suggests that if VWC is shifted forward, it will no longer be aligned with SWI, probably due to differences in how they react to environmental factors such as rainfall or the sensor technology. The cross-correlation analysis of the three locations shows that the relationship between VWC and the SWI varies across different locations, with the maximum lag differing from location to location. While two locations show a peak correlation at a lag of -3 days, indicating that VWC leads SWI by 3 days, the other one shows a maximum correlation at a lag of -2 days. This variability suggests that the response time of SWI to changes in VWC is influenced by location-specific factors such as soil properties and environmental factors. The strength of the correlation also varies across sites, ranging from modest at some locations to much higher at others. The presence of different lags highlights the need for localised modelling with consideration of local factors. Understanding these spatial differences can improve modelling

approaches and enhance predictions of soil moisture dynamics.

2.4.7 Exploring spatio-temporal patterns in Copernicus satellite soil moisture (via FRK).

This section explores spatio-temporal patterns in Copernicus satellite soil moisture data. Building on the earlier descriptive analysis, we move from simple plots and correlations to a model-based analysis that can reveal underlying structures.

It aims to characterise its spatial and temporal structure and how these patterns can be better captured and analysed in future modelling. By analysing the satellite data alone, we explore the underlying trends, variations, and dependencies that are not obvious from the raw data, to improve understanding of soil moisture dynamics within the satellite data.

Fixed Rank Kriging (FRK), a geostatistical model suited to large datasets has been applied (Cressie and Johannesson, 2008). FRK models broad spatial structure using basis functions while accounting for fine-scale variability through a residual component, thereby reducing computational cost. FRK helps identify spatial and temporal patterns in soil moisture satellite data and how they evolve across space and over time.

This approach enhances the resolution of the satellite estimates, which provides more reliable predictions to help understand inherent patterns. The remainder of this section summarises the FRK setup for the satellite data and presents the results of spatio-temporal patterns and uncertainty.

2.4.7.1 Methodology

We use Fixed Rank Kriging (FRK), a low-rank geostatistical framework that scales to large datasets by representing broad-scale structure via basis functions and modelling fine-scale variability through a residual component.

We adopt the spatial GLMM formulation:

$$\begin{aligned} Z_j \mid \mu_{Z_j}, \psi &\sim EF\left(\mu_{Z_j}, \psi\right), \quad j = 1, \ldots, m, \\ \boldsymbol{\mu}_Z = \mathbf{C}_Z \boldsymbol{\mu}, &\qquad g(\boldsymbol{\mu}) = Y, \\ Y &= \mathbf{T}\boldsymbol{\alpha} + \mathbf{S}\boldsymbol{\eta} + \boldsymbol{\xi}, \\ \boldsymbol{\eta} \sim N(\mathbf{0}, \mathbf{K}), &\qquad \boldsymbol{\xi} \sim N(\mathbf{0}, \sigma^2_\xi \mathbf{V}_\xi). \end{aligned} \tag{2.8}$$

Here, $\mathbf{T}$ contains BAU-level covariates (e.g., elevation, latitude, longitude), $\mathbf{S}$ evaluates a multi-resolution set of spatial (or spatio-temporal) basis functions, $\boldsymbol{\eta}$ are random coefficients with

covariance $\mathbf{K}$ (or precision $\mathbf{Q} = \mathbf{K}^{-1}$), and $\boldsymbol{\xi}$ captures fine-scale variation.

Basis Areal Units (BAUs) We discretise the domain into BAUs and aggregate to observation supports via $\mathbf{C}_Z$. This handles change-of-support and allows prediction on a regular grid.

Basis functions
We employ multi-resolution bisquare (or Gaussian) spatial bases; for spatio-temporal FRK, we take tensor products of spatial and temporal bases. Resolutions and counts are chosen to capture broad gradients and local features while keeping computation tractable (typically ≤ 3 spatial resolutions).

Estimation and prediction
Parameters are estimated with TMB using the Laplace approximation. We obtain predictions and their uncertainties on the BAU grid. In the Gaussian case, we summarise accuracy via RMSPE.

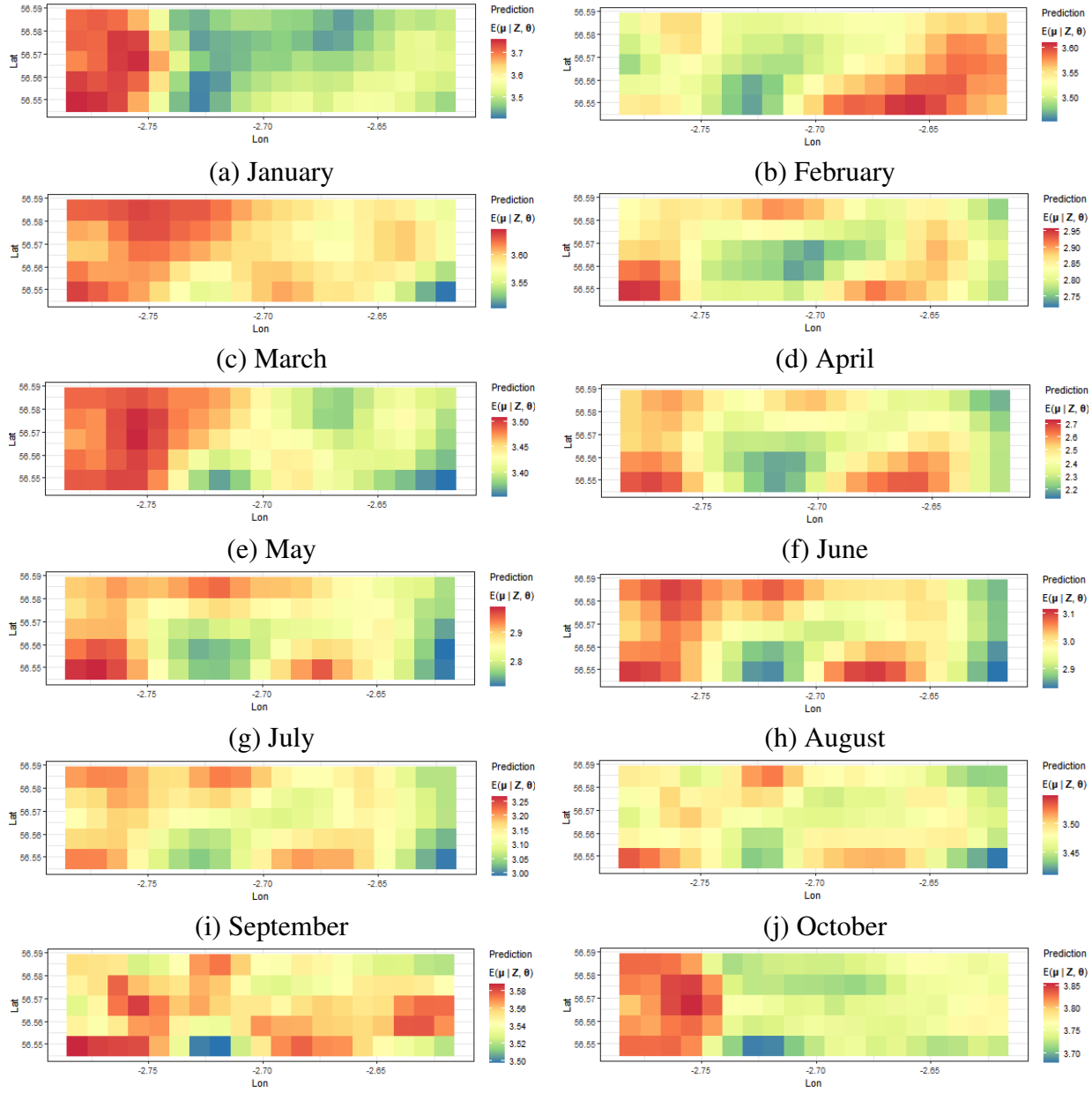
2.4.7.2 FRK setup

- Data: Copernicus Sentinel-1 soil moisture (SWI) over the Elliott Water catchment (Scotland), which includes 95 pixels for every day.
- Covariates: elevation, latitude, longitude (BAU level).
- Basis functions: FRK uses multi-resolution bisquare basis functions on a regular layout (3 spatial resolutions).
- Estimation: TMB (Laplace). Prediction grids: BAUs at 1 km.

2.4.7.3 Results

In order to explore whether there is any spatial pattern of the soil moisture in Elliot Water over time, the monthly averaged data are used to fit the model. The whole datasets (95 pixels) are separated into the training set (85) and the test set (10). The training set is used to fit the model, and the test data is used to test the accuracy of the prediction result. The residual plots are used to measure the accuracy of the prediction of the model.

The parameter estimation is carried out using the TMB method. The prediction is to make inferences based on the hidden process over the prediction regions D^P.



(a) January (b) February

(c) March (d) April

(e) May (f) June

(g) July (h) August

(i) September (j) October

(k) November (l) December

Figure 2.25: Monthly mean process from January to December. Each subfigure represents the mean process for a specific month, arranged chronologically from January to December.

Residuals through time

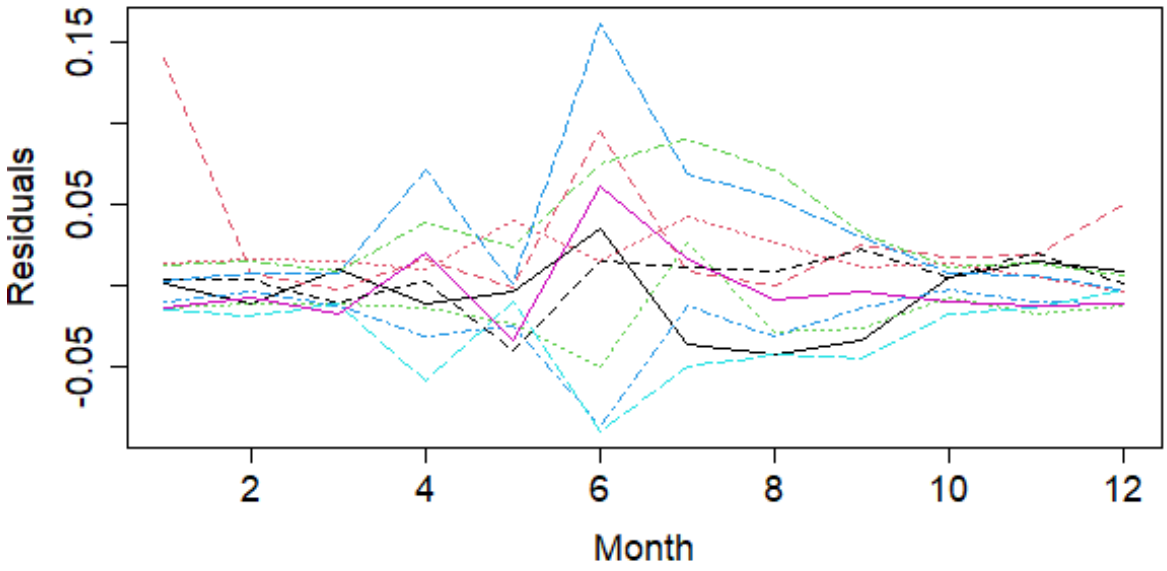


Figure 2.26: Residuals through time. High summer residuals and low autumn and winter residuals suggest reduced model flexibility in summer.

Figure 2.25 demonstrates the mean process from January to December, and Figure 2.26 shows that the model has more uncertainty during summer and less uncertainty during autumn and winter, which means that the model is less flexible in capturing the spatial variation in the summer months. More basic functions should be used in the FRK model to capture these spatial structures. From the prediction result, we can see some common spatial patterns from March to November. January and December show a similar pattern as well. February shows the opposite pattern to other months.

2.5 Conclusions

In this chapter, an exploratory analysis of soil moisture datasets is conducted to identify spatial-temporal patterns and relationships between variables that will inspire future studies. The analysis focused on two key soil moisture measurements: Volumetric water content (VWC), representing in situ soil moisture measurements, and the soil water index (SWI), satellite images derived from radar sensors. The well-maintained COSMOS In-situ soil moisture sensor is used here to understand the temporal patterns within the soil moisture data, as well as the relationships between VWC, air temperature, and precipitation. Temporal trends revealed a long-term trend in soil moisture, possibly related to strong seasonal fluctuations driven by precipitation and temperature changes. Spatial autocorrelation existed in spatial soil moisture patterns, suggesting

that adjacent locations share similar hydrological behaviours, while temporal autocorrelation highlighted persistence in moisture levels over the long term.

The relationship between the VWC and SWI provides important insights for further study:

- Pearson correlation between VWC and SWI ranged from 0.61 to 0.72 across different locations, indicating a moderate to strong linear relationship. However, it may inadequately capture the potential nonlinear relationship.
- Rolling correlation (15-day window) shows temporal dependencies, with coefficients fluctuating through time rather than staying constant. The difference between VWC and SWI can be caused by the differences in sensor technology(between the ground-based measurement or remote sensing), data processing algorithms or the reaction to environmental factors (e.g., SWI may lag because of integration of soil moisture over a large area compared to VWC direct measurement on a point location). The negative correlations suggest that the SWI may lag behind the rapid reaction captured by the ground-based in situ sensor, especially during rainfall events. The correlation shifts seem to align with the seasonal patterns, and the sudden drops may be linked to sensor errors or local environmental factors.
- The cross-correlation analysis shows that the relationship between VWC and the SWI varies across different locations, with the maximum lag differing from location to location. While two locations show a peak correlation at a lag of -3 days, indicating that VWC leads SWI by 3 days, the other one shows a maximum correlation at a lag of -2 days. This variability suggests that the response time of SWI to changes in VWC is influenced by location-specific factors such as soil properties and environmental factors. The presence of different lags highlights the need for localised modelling with consideration of local factors. Understanding these spatial differences can improve modelling approaches and enhance predictions of soil moisture dynamics.
- The FRK result summarises spatial patterns in Copernicus Sentinel-1 SWI over the Elliott Water catchment. We apply FRK to monthly-aggregated data on a 95-pixel grid, using multi-resolution bisquare basis function to capture broad structure with elevation, latitude, and longitude as BAU-level covariates. The mean maps smooth the raw satellite fields and highlight dominant trends and patterns across months. These summaries provide a description of large-scale soil-moisture structure and will guide the design and interpretation of the data fusion modelling in later chapters.

The exploratory analysis provides insights into the characteristics of soil moisture datasets used in the thesis, such as their variability, correlations, and spatial-temporal trends. These findings help shape the direction of the next chapter, which focuses on identifying and modelling the spatial-temporal patterns in soil moisture data obtained from different sources. By understanding

these patterns, we can better integrate soil moisture measurements from multiple datasets, coming up with more informed data fusion approaches.

Chapter 3

Spatio-temporal regression with misaligned covariates

3.1 Introduction

In Chapters 1 and 2, the data are introduced and a thorough exploratory analysis is carried out, which motivates the need for data fusion methods. This chapter tackles a key challenge: spatial misalignment between the response variable and covariates.

Spatial misalignment, which here refers to the response variable and the covariates being observed at different spatial locations, is a common challenge in many environmental research studies. (Scott, 2023) mentioned that it is very challenging to deal with the data fusion of misaligned data. Spatial regression models, commonly employed to investigate the relationship between response variables and covariates while considering spatial correlation (Cressie and Wikle, 2015), often assume that these variables are observed at the same locations. However, this is not always true in the real world. With the development of new technology, it has become increasingly common for response variables and covariates to be collected from different locations and data sources, such as environmental sensors gathering information from different collection points. As discussed in Chapter 2, Figure 3.1 shows the spatial distribution of the direct measurements of soil moisture (Volumetric Water Content, VWC), along with two variables that may have correlations with soil moisture: soil temperature and rainfall in the Elliot water catchment. In this context, rainfall is the misaligned covariate, whereas soil temperature is identified as the aligned covariate.

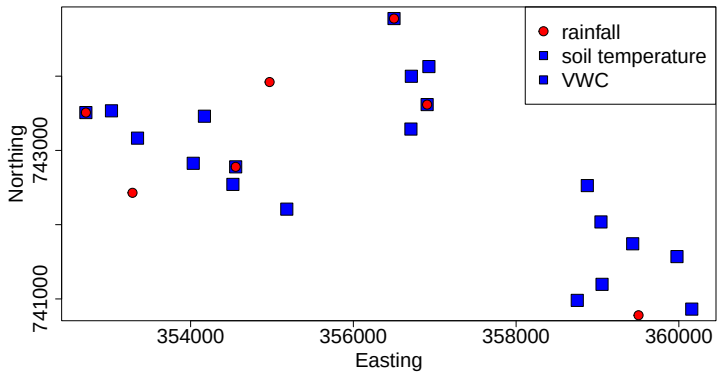


Figure 3.1: Locations for the response variable (VWC) and the aligned covariate (soil temperature) are represented by blue squares, while the misaligned covariate (rainfall) is represented by red circles.

However, most geostatistical approaches cannot accommodate misaligned covariates, requiring that both the covariates and the response variable be measured at the same locations and time points. In addition, these approaches treat all types of data as point data in the modelling, which ignores certain characteristics in other types of data. For example, in fields such as remote sensing, geospatial modelling, and climate modelling, geographic features or environmental variables are often represented in a regular raster, where each pixel contains a specific value, such as elevation or temperature, and each raster has spatial coordinates. In this chapter, a joint model is constructed to solve two problems: Include the misaligned covariate (in space and time) in the model to help with the prediction of the response variable and merge the grid data and point data to get better prediction results on the unobserved locations and parameter estimation. It is noted that only the SEPA data discussed in Section 2.1 are used in this chapter.

3.2 Recent work in the field

In this section, the relevant literature will be reviewed with a focus on the data fusion methods and the spatio-temporal misalignment issues within the INLA-SPDE framework. A comprehensive review of data fusion methods from a broader perspective can be found in Section 1.3.

3.2.1 Spatial and temporal misalignment

In environmental science, a typical way to deal with the misalignment problem is based on nearest-neighbour interpolation. This method predicts unobserved covariate values at the response variable location to be identical to the values at the closest measurement locations. However, this will lead to the underestimation of parameter variability within the model. An

alternative method is the krige-and-regress (KNR) method (Szpiro et al., 2011), which utilises Kriging to spatially align covariates with the response. It incorporates a Monte Carlo method to estimate the variance of the regression coefficient, accounting for additional variability introduced by the predicted covariate (Madsen et al., 2008). Szpiro et al. (2011) develop KNR and introduces three parametric bootstrap techniques for obtaining corrected variance estimators of the regression coefficient. More recently, a bootstrap approach for KNR has been applied under a survey sampling framework (Pouliot, 2023). It is noted that existing KNR literature typically considers a single misaligned covariate and assumes a linear relationship between the response and covariates, potentially imposing practical restrictions.

In the previous studies, most of the papers assume that the covariates are non-misaligned. Spatially misaligned data can be fused by a Bayesian hierarchical model, which assumes that each variable coming from the monitoring networks or satellites is a realisation of a continuously indexed spatial process (latent field) changing over time. Cameletti et al. (2013) develop a hierarchical spatio-temporal model for particulate matter (PM_{10}) concentration. This model includes a Gaussian Field (GF), impacted by a measurement error, and a state process characterised by a first-order autoregressive dynamic model and spatially correlated innovations. It considers a continuously indexed GF with Matérn covariance function as a discretely indexed random process to obtain spatio-temporal predictions and parameter estimation in a computationally efficient way. They implement this model to the point-referenced PM_{10} concentration to obtain the estimation of the parameters within the spatio-temporal model as well as the fine-resolution PM_{10} concentration map.

Krainski et al. (2018) shows a joint model that allows for the spatial misalignment between the response and the covariate, but only in spatial perspective, and does not incorporate the temporal dimension. Thus, to the best of our knowledge, numerous methods exist, but none explicitly address the misaligned covariates within the INLA-SPDE framework. Spatially misaligned data allows us to utilise all available data comprehensively. For instance, we may have only a limited number of sensors measuring soil moisture and related covariates at the same locations. However, we have additional sensors located at positions that do not align with the soil moisture sensors.

In a statistical framework, temporal misalignment refers to the discrepancy in timing between observations or measurements across different datasets or within a dataset over time (Box et al., 2015). As for the temporal misalignment, Zapata-Marin et al. (2023) propose a dynamic linear model (DLM) to predict unobserved fine-scale measurements from coarser-scale data, effectively handling the temporal aggregation issue in environmental data analysis. This work builds upon previous studies in temporal aggregation within DLMs, including Amemiya and Wu (1972) on autoregressive systems, Schmidt and Gamerman (1997) on the aggregated series following the same DLM class, Ferreira et al. (2006) on a multiscale model linking information across temporal scales, and Berrocal et al. (2010), who also contributed significantly to the field.

3.3 Methodology

3.3.1 Geostatistical model specification

The geostatistical model framework is defined as follows. The model assumes that there is a spatially continuous variable underlying all observations that can be modelled using a Gaussian random field process. Let D denote the subset that includes points that have real-number coordinates in a two-dimensional plane. The process is denoted by $S = \{S(\mathbf{x}) : \mathbf{x} \in D \subset \mathbb{R}^2\}$, has mean function $\mathbb{E}[S(\mathbf{x})] = 0$ and stationary covariance function $\mathrm{Cov}(S(\mathbf{x}), S(\mathbf{x}')) = \Sigma(\mathbf{x} - \mathbf{x}')$. Conditionally on $S(\mathbf{x})$, point data Y_i observed at a finite set of sites, say $\mathbf{x}_i \in D, i = 1, 2, \ldots, I$, are mutually independent with

$$Y_{pi}(\mathbf{x}_i) \mid S(\mathbf{x}_i) \sim N\left(\mu(\mathbf{x}_i) + S(\mathbf{x}_i), \tau^2\right),$$

where $\mu(\mathbf{x}_i)$ represents the large scale structure.

3.3.2 The framework of Integrated Nested Laplace Approximation (INLA)

INLA focuses on models that can be expressed as latent Gaussian Markov random fields (GMRF). The INLA framework can be described as follows: $\mathbf{y} = (y_1, \ldots, y_n)$ is a vector of observed variables whose distribution is in the exponential family, and the mean μ_i (for observation y_i) is linked to the linear predictor η_i using an appropriate link function. The linear predictor can include fixed effects and different random effects. $\mathbf{X}$ denotes the matrix of all latent effects which include the linear predictor, coefficients, and the distribution of the vector of latent effects is assumed to be Gaussian Markov random field (GMRF) with a zero mean and precision matrix $\mathbf{Q}(\boldsymbol{\theta}_2)$, with $\boldsymbol{\theta}_2$ a vector of hyperparameters. The distribution of $\mathbf{y}$ will depend on some vector of hyperparameters $\boldsymbol{\theta}_1$. The vectors of all hyperparameters in the model will be denoted by $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

Observations are assumed to be independent given the latent effects and the hyperparameters, which means the likelihood can be written as

$$\pi(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}) = \prod_{i \in \mathscr{I}} \pi(y_i \mid \eta_i, \boldsymbol{\theta}),$$

where set $\mathscr{I}$ contains indices for all observed values of $\mathbf{y}$.

The joint posterior distribution of the effects and hyperparameters can be expressed as:

$$\pi(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta})\pi(\mathbf{x} \mid \boldsymbol{\theta}) \prod_{i \in \mathscr{I}} \pi\left(y_i \mid x_i, \boldsymbol{\theta}\right)$$
$$\propto \pi(\boldsymbol{\theta})|\mathbf{Q}(\boldsymbol{\theta})|^{1/2} \exp\left\{-\frac{1}{2}\mathbf{x}^{\top}\mathbf{Q}(\boldsymbol{\theta})\mathbf{x} + \sum_{i \in \mathscr{I}} \log\left(\pi\left(y_i \mid x_i, \boldsymbol{\theta}\right)\right)\right\},$$

where $\mathbf{Q}(\boldsymbol{\theta})$ to represent the precision matrix of the latent effects.

The marginal distributions for the latent effects and hyperparameters can be calculated from:

$$\boldsymbol{\pi}\left(x_i \mid \mathbf{y}\right) = \int \boldsymbol{\pi}\left(x_i \mid \boldsymbol{\theta}, \mathbf{y}\right) \pi(\boldsymbol{\theta} \mid \mathbf{y}) d\theta,$$

and

$$\boldsymbol{\pi}\left(\theta_j \mid \mathbf{y}\right) = \int \boldsymbol{\pi}(\boldsymbol{\theta} \mid \mathbf{y}) d\theta_{-j}.$$

Since both of the marginal distributions include integration over the space of the hyperparameters, and the dimension of $\boldsymbol{\theta}$ depends on the number of observations, which means that numerical integration is difficult for high dimensional data, a good approximation of the joint posterior distribution of the hyperparameters is required. Rue et al. (2009) approximate $\boldsymbol{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$, denoted by $\tilde{\boldsymbol{\pi}}(\boldsymbol{\theta} \mid \mathbf{y})$, which is achieved by using the computational properties of GMRF and the Laplace approximation for multidimensional integration, and use this to approximate the posterior marginal of the latent parameter x_i as:

$$\tilde{\boldsymbol{\pi}}\left(x_i \mid \mathbf{y}\right) = \sum_k \tilde{\boldsymbol{\pi}}\left(x_i \mid \theta_k, \mathbf{y}\right) \times \tilde{\boldsymbol{\pi}}\left(\theta_k \mid \mathbf{y}\right) \times \Delta_k,$$

where Δ_k are the weights associated with a vector of values θ_k of the hyperparameters in a grid.

3.3.3 SPDE approach

Lindgren et al. (2011) consider a stochastic partial differential equation (SPDE) whose solution is a Gaussian field (GF) with Matérn correlation and proposes a new approach to represent a GF with Matérn covariance, as a GMRF, by representing a solution of SPDE using the finite element method. The benefit is that the GMRF representation of the GF, which can be computed explicitly, provides a sparse representation of the spatial effect through a sparse precision matrix, which enables the nice computational properties of the GMRF, which can then be implemented in the INLA approach. To be specific, GMRF is a discrete approximation of a Gaussian field. It is obtained by discretising the continuous domain into a grid or lattice of points. In a GMRF, the values at each grid point are assumed to be conditionally independent of all other points, given their neighbouring points. This conditional independence property is often represented using a sparse precision matrix (also known as an inverse covariance matrix), where nonzero entries

indicate dependencies between neighbouring points. GMRFs provide a computationally efficient way to model and analyse large spatial datasets.

The linear fractional SPDE can be defined as:

$$\left(\kappa^2 - \Delta\right)^{\alpha/2} s(\mathbf{x}) = \mathbf{W}(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d, \quad \alpha = \nu + d/2, \quad \kappa > 0, \quad \nu > 0, \tag{3.1}$$

where Δ is the Laplacian operator: $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$ and $\mathbf{W}(\mathbf{x})$ denotes a spatial white noise Gaussian stochastic process with unit variance.
Given n observations $y_i, i = 1, \ldots, n$, at locations $\mathbf{x}_i$, the following model can be defined:

$$\mathbf{y} \mid \beta_0, \mathbf{s}, \sigma_e^2 \sim N\left(\beta_0 + \mathbf{A}\mathbf{s}, \sigma_e^2\right), \tag{3.2}$$

$$\mathbf{s} \sim GF(0, \Sigma), \tag{3.3}$$

where β_0 is the intercept, $\mathbf{A}$ is the projection matrix and $\mathbf{s}$ is a spatial Gaussian random field. Note that the projection matrix $\mathbf{A}$ links the spatial Gaussian random field (defined using the mesh nodes, which are similar to the integration points on a numeric integration algorithm) to the locations of observed data.

3.3.3.1 Gaussian random field process

A Gaussian field (GF) process can be denoted by $S(\mathbf{x})$, where $\mathbf{x}$ is any location in a study area $\mathbf{D}$. $S(\mathbf{x})$ is a stochastic process, with $\mathbf{X} \in \mathbf{D}$, where $\mathbf{D} \subset \mathbb{R}^d$. For example, $\mathbf{D}$ is a domain and data have been collected at geographical locations, over $d = 2$ dimensions within this domain. The continuously indexed GF is assumed to be continuous over space and implies that it is possible to collect data at any finite set of locations within the domain. To complete the specification of the distribution of $S(\mathbf{x})$, it is necessary to define its mean and covariance. A very simple way to define a correlation function is based only on the Euclidean distance between locations, which assumes that when two pairs of points are equally distant from each other, they will exhibit an equivalent level of correlation. Matérn covariance is another widely used way to define the correlation function, and the details are in Equation (3.4).

In many scenarios, it is commonly assumed that there exists an underlying GF that cannot be directly observed. Instead, observations are data with a measurement error e_i,

$$y(\mathbf{x}_i) = S(\mathbf{x}_i) + e_i,$$

where e_i is independent of e_j for all $i \neq j$ and e_i follows a Gaussian distribution with zero mean and variance σ_e^2. The covariance of the marginal distribution of $y(\mathbf{x})$ at a finite number of locations is $\Sigma_y = \Sigma + \sigma_e^2 \mathbf{I}$.

3.3.3.2 The Matérn covariance

The Matérn covariance is widely used in various scientific fields to define the covariance function Σ, and the reason it is used here is that GF $s(\mathbf{x})$ with the Matérn covariance are a solution to the linear fractional SPDE shown in Equation (3.1).

For two locations $\mathbf{x}_i$ and $\mathbf{x}_j$, the stationary and isotropic Matérn correlation function is defined as:

$$\text{Cor}_M\left(S\left(\mathbf{x}_i\right), S\left(\mathbf{x}_j\right)\right) = \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\kappa\left\|\mathbf{x}_i - \mathbf{x}_j\right\|\right)^{\nu} K_{\nu}\left(\kappa\left\|\mathbf{x}_i - \mathbf{x}_j\right\|\right), \tag{3.4}$$

where $\|.\|$ denotes the Euclidean distance and K_{ν} is the modified Bessel function of the second kind and ν is the order. To be specific, the modified Bessel function of the second kind is the function $K_n(x)$, which is one of the solutions to the modified Bessel differential equation. κ is a scaling parameter, which can also be interpreted as a range parameter ρ, representing the Euclidean distance at which x_i and x_j become almost independent. The empirically derived definition $\rho = \sqrt{0.8\nu}/\kappa$, corresponds to correlation near 0.1 at the distance ρ, for all ν.

The Matérn covariance function is $\sigma_u^2 \text{Cor}_M\left(S\left(\mathbf{x}_i\right), S\left(\mathbf{x}_j\right)\right)$, where σ_u^2 is the marginal variance of the process and is defined as:

$$\sigma_u^2 = \frac{\Gamma(\nu)}{\Gamma(\nu + d/2)(4\pi)^{d/2}\kappa^{2\nu}}$$

3.3.3.3 Basis functions

The domain D can be divided into a set of non-intersecting triangles, where any two triangles meet in at most a common edge or corner. The corners are named vertices. The solution for the SPDE will depend on the basis functions used. The basis functions used by Lindgren et al. (2011) to construct the solution $s(\mathbf{x})$ in the SPDE is defined as:

$$s(\mathbf{x}) = \sum_{k=1}^{m} \psi_k(\mathbf{x}) w_k,$$

where ψ_k is the basis function and w_k is the Gaussian-distributed weights, and m is the number of vertices in the triangulation.

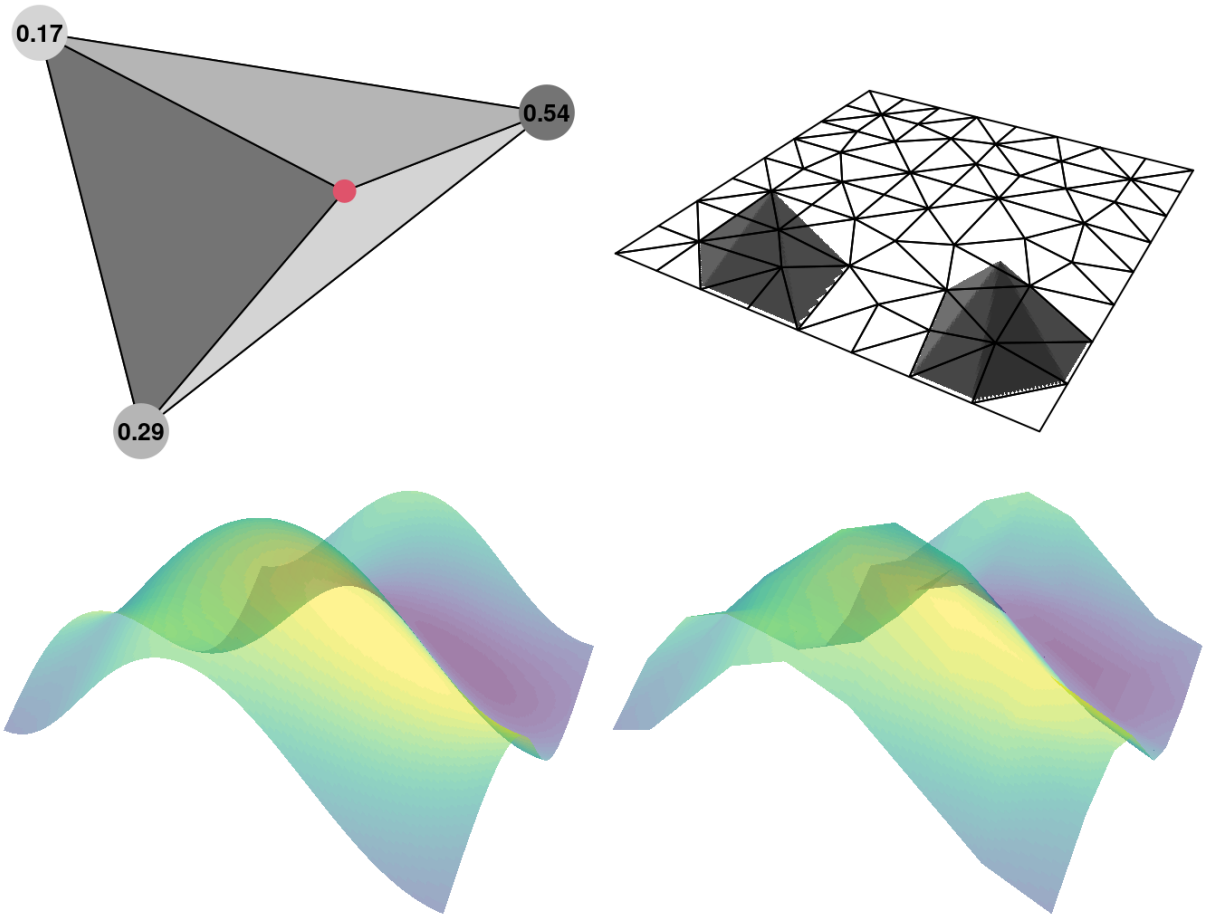


Figure 3.2: A triangle and the scenario exemplify the use of barycentric coordinates for the point in red (top left). All the triangles and the basis function for two of them (top right). A true field for illustration (bottom left) and its approximated version (bottom right). Source: Krainski et al. (2018).

The large triangle in the upper left corner of Figure 3.2, contains a red dot and the trio of smaller triangles generated by connecting this point with the large triangle's vertices. The numerical values at the vertices of the large triangle represent the ratio of the area of the adjacent small triangle (not sharing that vertex) to the total area of the large triangle. Thus, the sum of these three values equals one. These values correspond to the evaluation of the basis function at the red dot, based on its position relative to the vertices of the large triangle. They serve as coefficients in the approximation, influencing the function's value at each vertex of the large triangle.

The interpretation of the equation depends on the choice of the basis function, for example, if ψ_k is 1 at vertex k and zero at all other vertices. The weights determine the values of the field at the vertices, and the values inside the triangles are linearly interpolated.

3.3.3.4 Projection matrix

When working with spatial data collected at a set of locations, an objective is to predict the spatial model on a fine-scale grid to generate high-resolution maps. The projection matrix **A**, which is used to get the map of the random field on a fine grid, can be used for interpolating the posterior mean of the random field.

In matrix form, this concept is related to the projector matrix A. In Figure 3.2, for a point situated within a triangle, the respective row in A contains three non-zero entries. If the point lies on an edge, two entries are non-zero, and for a point that lies at a vertex, a single non-zero entry exists, which is one. The dimension of the projection matrix equals the number of data locations times the number of vertices in the mesh, and each point location is either inside one of the triangles or at a vertex, so there are no more than three elements in each row that are non-zero.

3.3.4 Penalised Complexity priors

Simpson et al. (2017) develops the Penalised Complexity priors (PC priors), which set the prior of standard deviation σ of the latent field by defining the parameters μ and α. The definition of the PC priors is as follows:

$$\text{Prob}(\sigma > u) = \alpha, u > 0, 0 < \alpha < 1,$$

which means that for this latent field, the probability of the standard deviation being higher than μ is α %. To be specific, the PC prior to the precision τ has a density

$$\pi(\tau) = \frac{\lambda}{2}\tau^{-3/2}\exp\left(-\lambda\tau^{-1/2}\right), \quad \tau > 0$$

for $\lambda > 0$ where

$$\lambda = -\frac{\ln(\alpha)}{u}$$

and (u, α) are the parameters to this prior. The interpretation of (u, α) is that

$$\text{Prob}(\sigma > u) = \alpha, \quad u > 0, \quad 0 < \alpha < 1,$$

where the standard deviation is $\sigma = 1/\sqrt{\tau}$. The density, cumulative distribution function, and quantile function. R-INLA uses the log-precision rather than the precision, and the corresponding PC prior to the log-precision x has a density

$$\pi(x) = \frac{\lambda}{2}\exp\left(-\lambda\exp\left(-\frac{x}{2}\right) - \frac{x}{2}\right).$$

The joint PC prior density for the spatial range, ρ, and the marginal standard deviation, σ, is given by:

$$\pi(\rho,\sigma) = \frac{d\lambda_\rho}{2}\rho^{-1-d/2}\exp\left(-\lambda_\rho\rho^{-d/2}\right)\lambda_\sigma\exp\left(-\lambda_\sigma\sigma\right)$$

where λ_ρ and λ_σ are hyperparameters that must be determined according to the information. In INLA, the practical approach to setting these hyperparameters involves indirectly specifying them through:

$$P(\rho < \rho_0) = p_\rho$$

and

$$P(\sigma > \sigma_0) = p_\sigma$$

The lower tail quantile and probability for the range (ρ_0 and p_ρ) and the upper tail quantile and probability for the standard deviation (σ_0 and p_σ) need to be specified. This allows the user to control the priors of the parameters by providing knowledge of the scale of the problem. What is a reasonable upper magnitude for the spatial effect, and what is a reasonable lower scale at which the spatial effect can operate? The shape of the prior was derived through a construction that shrinks the spatial effect towards a base model of no spatial effect in the sense of distance measured by Kullback-Leibler divergence.

The prior is constructed in two steps, under the idea that having a spatial field is an extension of not having a spatial field. First, a spatially constant random effect ($\rho = \infty$) with finite variance is more complex than not having a random effect ($\sigma = 0$). Second, a spatial field with spatial variation ($\rho < \infty$) is more complex than the random effect with no spatial variation. Each of these extensions is shrunk towards the simpler model and, as a result, we shrink the spatial field towards the base model of no spatial variation and zero variance ($\rho = \infty$ and $\sigma = 0$). The details behind the construction of the prior are presented in Fuglstad et al. (2019) and are based on the PC prior framework (Simpson et al., 2017).

3.3.5 Cross-validation

Cross-validation is a statistical method used to estimate the performance of models. It includes separating the original dataset into a training set to train the model and a test set to evaluate its performance. This helps assess how the model will be generalised to an independent dataset. It is particularly useful in scenarios where the objective is to predict the value of a new data point not seen by the model. Two model-fitting measurements are used for evaluating the goodness of the model, and they are based on the predictive distribution. There are also some indices based on deviance, such as DIC, but they can only be used for model comparison when all the models are fitted to the same dataset.

3.3.5.1 Deviance information criterion (DIC)

The deviance information criterion (DIC), introduced by Spiegelhalter et al. (2002), is a commonly used model-fitting measurement for Bayesian models. It extends the Akaike information criterion (AIC) and is specifically designed to compare Bayesian models. DIC has two components: one assesses the model fit, and the other evaluates its complexity. The measure of model fit is represented by the posterior expectation of the deviance $D(\boldsymbol{\theta}) = -2log(p(\mathbf{y}|\boldsymbol{\theta}))$, while the model's complexity is quantified by the effective number of parameters:

$$p_{\mathrm{D}} = \mathrm{E}_{\boldsymbol{\theta}|y}(\mathrm{D}(\boldsymbol{\theta})) - \mathrm{D}\left(\mathrm{E}_{\boldsymbol{\theta}|y}(\boldsymbol{\theta})\right) = \overline{\mathrm{D}} - \mathrm{D}(\overline{\boldsymbol{\theta}}),$$

and the DIC is

$$\mathrm{DIC} = \overline{\mathrm{D}} + p_{\mathrm{D}},$$

where $\overline{\mathrm{D}}$ is the posterior mean of the deviance, $\mathrm{D}(\overline{\boldsymbol{\theta}})$ is the deviance of the posterior mean of the parameters and p_{D} is the effective number of parameters. In INLA, instead of assessing the deviance based on the posterior mean of all parameters, it is evaluated using the posterior mean of the latent field $\boldsymbol{\theta}$ and the posterior mode of the hyperparameters ψ. This choice is made because the posterior marginals of certain hyperparameters, particularly precisions, may exhibit significant skewness. Consequently, the posterior expectation may not accurately reflect the distribution, and the mode is considered a more suitable representation.

3.3.5.2 Conditional predictive ordinate (CPO)

The conditional predictive ordinate (CPO) allows assessing the model's predictive performance for individual observations by quantifying the likelihood of observing a specific data point given the rest of the data. As an index for detecting surprising observations y_i within a model and therefore checking the model fit, the CPO for each observation can be computed. To be more precise, this predictive quantity is given by:

$$CPO_i = \pi(y_i|\mathbf{y}_{-i})$$

To be specific, CPO calculates the predictive probability of observing a specific data point y_i within a given model, conditioned on the remaining data $\mathbf{y}_{-i}$. It can be interpreted as the likelihood of observing the specific data point y_i if the model were true and the remaining data $\mathbf{y}_{-i}$ were fixed. A low CPO value for a specific observation suggests that the model has a poor prediction on that observation, indicating that it might be an outlier or an unusual data point. Conversely, a high CPO value indicates a good fit and predictability for that observation within the model. However, CPO values need to be adjusted based on the Gaussian field's level to make them comparable. David and Johnson (1948) propose a calibration procedure called the probability integral transform (PIT) that can be used for this purpose.

3.3.5.3 Probability integral transform (PIT)

PIT is a tool for assessing the adequacy of a single model, which is defined as:

$$PIT_i = Pr(y_{\text{new}_i} \leq y_i | \mathbf{y}_{-i}),$$

where $\mathbf{y}_{-i}$ refers to the observation vector with the ith observation removed. The value obtained from the predictive cumulative distribution function (CDF) at the observation y_i corresponds to this omitted component. This procedure is calculated in a cross-validation mode, in which each step of the validation process involves calculating the posterior predictive distribution by leaving out one observation at a time.

A uniformly distributed histogram of the PITs indicates a good model and a lower possibility for forecast failures. U-shaped histogram indicates under-dispersed predictive distributions, while inverse-U-shaped histograms point to over-dispersion, and skewed histograms occur when central tendencies are biased.

3.4 Simulation study

3.4.1 Aim

In many applications, real data tends to be complex, noisy, and hard to explain. Particularly in environmental applications, the monitoring network used for data collection is often extremely sparse. The severe scarcity limits the number of locations and time points in the real dataset, making it challenging to model spatial correlations due to the small number of available locations, even if the variable of interest is genuinely continuous across space and through time. Therefore, a simulation study is employed to assess the effectiveness of the model (3.6) developed in this situation. The simulation study is designed to determine the minimum number of points required for parameter estimation and to understand how the features of the latent fields impact the parameter estimation and prediction ability.

Previous studies have conducted simulations similar to the situation here. For example, Moraga et al. (2017) simulate four surfaces with different degrees of roughness and marginal standard deviation in the Matérn field. They apply the geostatistical model presented in this paper to real data, which consists of 155 grid cells (10km) and 14 monitoring sites scattered sparsely throughout the same region, both measuring the fine particulate air pollutant $PM_{2.5}$. The study revealed that the method doesn't precisely recover the true values of the parameters used in the simulation study, but it did show that the 95% credible intervals contained the true values for most of the parameters within the model. However, the real data from Elliot water include both aligned covariates and misaligned covariates, so the misaligned covariates needed to be adapted

into the model framework proposed in this paper.

3.4.2 Model framework and parametrisation

The model framework is an extension of the base model in Equation (3.5) presented in Chapter 3 of the work by Krainski et al. (2018), while the base model incorporates just a single misaligned covariate.

The base model can be separated into fixed effects, random effects, and measurement error variance, which can be expressed as follows:

$$\begin{aligned} y_1(\mathbf{s}) &= \mu_1(\mathbf{s}) + e_1(\mathbf{s}) \\ y_2(\mathbf{s}) &= \alpha_1 + \beta_1(\mu_1(\mathbf{s})) + \mu_2(\mathbf{s}) + e_2(\mathbf{s}), \end{aligned} \tag{3.5}$$

where the α_1 is the intercept, $\mu_k(\mathbf{s})$ are spatial effects, β_1 is the scaling parameter for the spatial effect and $e_k(\mathbf{s}) \sim N(0, \sigma^2_{ek})$ are uncorrelated error terms defined by a Gaussian white-noise process, with $k = 1, 2$.

3.4.2.1 Joint model with one misaligned covariate, one non-misaligned covariate, and fixed effects

Equation (3.6) expands Equation (3.5) by introducing fixed effects and multiple covariates. The spatial-only joint model is defined considering the following equations:

$$\begin{aligned} y_1(\mathbf{s}^*) &= \alpha_1 + \mu_1(\mathbf{s}^*) + e_1(\mathbf{s}^*) \\ y_2(\mathbf{s}) &= \alpha_2 + \mu_2(\mathbf{s}) + e_2(\mathbf{s}) \\ y_3(\mathbf{s}) &= \alpha_3 + \beta_3 x(\mathbf{s}) + \beta_1(\alpha_1 + \mu_1(\mathbf{s})) + \beta_2(\alpha_2 + \mu_2(\mathbf{s})) + \mu_3(\mathbf{s}) + e_3(\mathbf{s}), \end{aligned} \tag{3.6}$$

where $y_k(\mathbf{s})$ denotes the realization of the spatial process $Y(\cdot)$ which represents the variables measured at location s. The α_k are the intercepts, $\mu_k(\mathbf{s})$ are spatial effects, β_1 and β_2 are scaling parameters for some of the spatial effects, β_3 is the scaling parameter of the fixed effect and $e_k(\mathbf{s}) \sim N(0, \sigma^2_{ek})$ are uncorrelated error terms defined by a Gaussian white-noise process, with $k = 1, 2, 3$, and it is spatially uncorrected. Further, $x(\mathbf{s})$ is the fixed effect. s^* here denotes that the data of the specific variable is collected at non-coinciding locations with other variables, and s here denotes that the data of the specific variable are collected at the same locations as other variables.

Equation (3.7) builds upon Equation (3.6) by incorporating spatio-temporal random effects. The temporal process is assumed to be AR(1) in the simulation. The model can be used to predict soil moisture by including both aligned and misaligned covariates in the model. The spatio-temporal joint model is defined considering the following equations:

$$
\begin{aligned}
y_1(\mathbf{s}^*,\mathbf{t}^*) &= \alpha_1 + z_1(\mathbf{s}^*,\mathbf{t}^*) + e_1(\mathbf{s}^*,\mathbf{t}^*) \\
y_2(\mathbf{s},\mathbf{t}) &= \alpha_2 + z_2(\mathbf{s},\mathbf{t}) + e_2(\mathbf{s},\mathbf{t}) \\
y_3(\mathbf{s},\mathbf{t}) &= \alpha_3 + \beta_3 x(\mathbf{s},\mathbf{t}) + \beta_1(\alpha_1 + z_1(\mathbf{s},\mathbf{t})) + \beta_2(\alpha_2 + z_2(\mathbf{s},\mathbf{t})) + z_3(\mathbf{s},\mathbf{t}) + e_3(\mathbf{s},\mathbf{t}),
\end{aligned}
\tag{3.7}
$$

where $y_k(\mathbf{s},\mathbf{t})$ denotes the realization of the spatio-temporal process $Y(\cdot,\cdot)$ which represents the variables measured at location located at location s and time point t. The α_k are the intercepts, $z_k(\mathbf{s},\mathbf{t})$ are space-time effects, β_1 and β_2 are scaling parameters for space-time effects, β_3 is the scaling parameter of the trend covariate and $e_k(\mathbf{s},\mathbf{t}) \sim N(0,\sigma^2_{ek})$ are uncorrelated error terms defined by a Gaussian white-noise process, with $k = 1,2,3$, and it is spatially and serially uncorrected. Further, $x(\mathbf{s})$ is the fixed effect, s^* and t^* here denote that the data of the specific variable is collected at non-coinciding locations and time points with other variables, and s and t here denote that the data of the specific variable is collected at the same locations with other variables.

3.4.3 Spatio-only model

The simulation study of the spatial-only model focuses on the model defined in Equation (3.6), and the true values of the parameters defined within the model are shown in Table 3.1 and Equation (3.2) for each scenario. The choice of parameters within the model will be justified in the following sections. Notably, in the simulation study, it is assumed that the misaligned covariate is available for both point data and grid data, but in reality, it is only available for the point data (sensor data), not the grid data (satellite data), which means satellite data on rainfall and soil temperature for this area in the real dataset is not available. The simulation still assumes that misaligned covariates are available to test the models performance under ideal conditions. This allows us to evaluate the model with both aligned and misaligned covariates, despite the practical limitations in the real data.

3.4.3.1 Parameters for data simulation

The values of spatial parameters for the random fields ($\mu_1(\mathbf{s})$, $\mu_2(\mathbf{s})$ and $\mu_3(\mathbf{s})$), such as the spatial variances of the Matérn covariance function, are chosen based on previous work (Blangiardo and Cameletti, 2015). Similarly, the values of the scale parameter β_1 and β_2, intercepts α_k, and error term e_k are chosen in accordance with the previous work (Moraga et al., 2017). The value of the scaling parameter β_3 for the fixed effect, which describes the relationship between the covariate x_3 and the response variable y, is chosen based on the observed relationship between elevation and VWC in the real dataset. To be clear, the parameters used in the previous work are reused in this simulation study, and for the fixed effect, which was not included in the previous study, the value which mimics the relationship in the real data is used here. The decision to model a covariate as either a fixed effect or a random effect depends on its availability at the predicted

target locations. For instance, elevation is accessible across the entire domain, allowing it to be treated as a fixed effect. Conversely, variables like rainfall and soil temperature, which have limited observations, are treated as random effects due to their limited availability.

The joint model specified in Equation (3.6) can be fitted using the INLA approach. The details of this approach can be found in the Section 3.4.3.6. Let $\boldsymbol{\theta}$ denote the vector of the hyperparameters, then

$$\boldsymbol{\theta} = \left(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \boldsymbol{\rho}_3, \boldsymbol{\sigma}_1^2, \boldsymbol{\sigma}_2^2, \boldsymbol{\sigma}_3^2, \boldsymbol{\sigma}_{e_1}^2, \boldsymbol{\sigma}_{e_2}^2, \boldsymbol{\sigma}_{e_3}^2\right).$$

The posterior marginals of $\boldsymbol{\theta}$ is approximated as $\boldsymbol{\pi}\left(\theta_j \mid \mathbf{y}\right) = \int \boldsymbol{\pi}(\boldsymbol{\theta} \mid \mathbf{y}) d\theta_{-j}$.

3.4.3.2 Measurement error model

The measurement error model is typically used to account for the uncertainty of covariate measurements. The simple story is that if a response variable $y(s)$ depends on a covariate $x(s)$ and both of them vary over space, $x(s)$ is assumed to be observed with error. There are two widely used measurement error models: the classical measurement error model (MEC) and the Berkson measurement error model (MEB). The difference between these two models is the assumption of the dependence of the measurement error. To be specific, the MEC is defined as: $w = x + \varepsilon$ and MEB is defined as $x = w + \varepsilon$, where w is the observed value for covariate x and ε is the error. MEC assumes that ε is independent of x while MEB assumes that ε is independent of w and both of them are non-differential (Muff et al., 2015). The impacts of classical and Berkson errors on these estimates are opposite. For example, the MEC overestimates the upper percentile of x and underestimates the lower percentile of x, while MEB underestimates the upper percentile of x and overestimates the lower percentile of x. In a simple story, w is more variable than the true covariate x in the classical model, whereas the opposite is true in the Berkson case.

Ignoring the measurement error will cause biases in the parameter estimation and mask important features of the data. The error variance and the error model needed to be specified correctly to ensure the error correction. The model in Equation (3.6) extends the base model by incorporating a classical ME into the framework of the latent Gaussian models.

3.4.3.3 Simulation strategy

In the real case, the sensors to collect direct measurement indices of soil moisture (VWC) and other potential variables related to soil moisture (rainfall, soil temperature, air temperature) might be employed at different locations. The simulation data will mimic the data as far as possible, and misaligned covariates, aligned covariates, and response variables are simulated as follows:

1. The spatial process $\mu(s)$ is simulated by generating independent random field realisations from a Matérn Gaussian random field. The behaviour of the Matérn Gaussian field is

controlled through three parameters: range (ρ), marginal variance (σ), and smoothness (v). The Matérn correlation function is used to generate the Matérn GRF.

2. The trend covariate $x(s)$, which represents the geological characteristics of the area, is derived from a surface where values exhibit an increasing pattern from the southwest to the northeast (from 0 to 3.5) across the study area. Let the coordinates of $y(s_i)$ be denoted by Easting(s_i) and Northing(s_i), then the trend is formulated as $x(s_i) = 0.2 * Easting(s_i) + 0.3 * Northing(s_i)$. Additionally, the geographic trend parameter β_3 is defined as -0.2. This may correspond to a surface indicating variations in variables such as soil moisture or other environmental covariates that are associated with changes in latitude and longitude. Figure 3.3 shows the surface of the trend covariate $x(s)$.

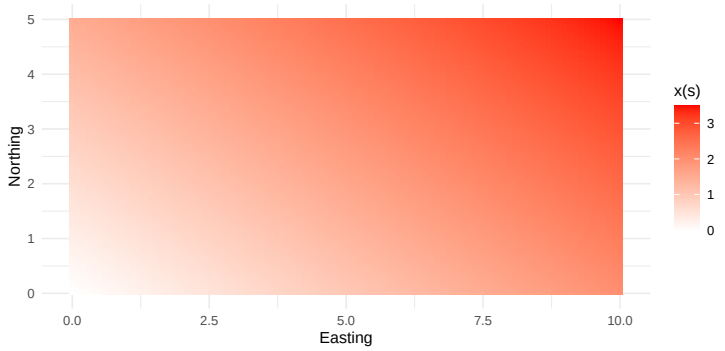


Figure 3.3: Surface of the trend covariate $x(s)$.

3. The uncorrected error terms are generated from a Gaussian white-noise process: $N(0, \sigma_{ek})$.

4. Then the covariates and the response variables are generated by combining the previously constructed terms based on Equation (3.6).

5. The test set includes 20 locations for the response variable y_3 in each scenario, and they are randomly generated in each simulation. It is noticed that the locations in the test set are different for each simulation.

There are 8 scenarios in the simulation study, which can be divided into two groups according to the two aims of the simulation study:

- Evaluate the impacts of spatial parameter ρ and marginal standard deviation σ of the Matérn field on the accuracy of parameter estimation.
- Evaluate the impacts of the number of points on the accuracy of parameter estimation.

All the intercepts $(\alpha_1, \alpha_2, \alpha_3)$, scaling parameters $(\beta_1, \beta_2, \beta_3)$, and precisions of the errors $(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \frac{1}{\sigma_3^2})$ are all fixed for all scenarios (same as the values in Scenario1). Scenario 1 and Scenario 2 are designed to evaluate the impacts of the spatial parameter ρ of the Matérn field on the accuracy of parameter estimation. Scenario 2 has a large range, and Scenario 3 has a small range, and the difference between these two simulation surfaces is shown in Table 3.1. Scenario 1 and Scenario 3 are designed to evaluate the impacts of the marginal standard deviation σ of the Matérn field on the accuracy of parameter estimation. Scenario3 has a large σ and Scenario 4 has a small σ, and the difference between these two simulations surfaces is shown in Table 3.1. Scenarios 1 and 4,5,6,7,8 are designed to evaluate the impacts of the number of points on the accuracy of parameter estimation. Table 3.2 details the number of locations for each variable across the scenarios.

Table 3.1: Parameters of the simulated surfaces within scenarios used to assess the impacts of varying Matérn field range and marginal standard deviation.

	Scenario1 (10,22,22)	Scenario2 (10,22,22)	Scenario3 (10,22,22)
α_1	0.5	0.5	0.5
α_2	0.8	0.8	0.8
α_3	1	1	1
β_1	-0.3	-0.3	-0.3
β_2	-0.4	-0.4	-0.4
β_3	-0.2	-0.2	-0.2
ρ_1	4	2	4
ρ_2	3	1.5	3
ρ_3	2	1	2
σ_1	1	1	4
σ_2	0.5	0.5	2
σ_3	0.3	0.3	1.2
$\sigma_{e_1}^2$	0.36	0.36	0.36
$\sigma_{e_2}^2$	0.25	0.25	0.25
$\sigma_{e_3}^2$	0.16	0.16	0.16

Table 3.1 displays the parameters for the simulated surfaces across different scenarios. The true values of the number of locations in the real data are $n_1 = 10, n_2 = 22$, and $n_3 = 22$, respectively. Specifically, scenario 3 represents a highly smooth surface with the same range parameter as scenario one but greater variance. Scenario 2, a rough surface, has a smaller range and equal variance compared to scenario 1, making it significantly more heterogeneous and hence more challenging to estimate.

Table 3.2: Parameters of the simulated surfaces within scenarios used to assess the impacts of varying numbers of locations.

	Number of locations
Scenario1	(10,22,22)
Scenario4	(20,44,44)
Scenario5	(40,88,88)
Scenario6	(80,176,176)
Scenario7	(100,220,220)
Scenario8	(200,440,440)

For each scenario, 100 independent replications are performed to evaluate the performance of the joint model. For each estimated posterior marginal distribution, 200 random values were simulated to compute posterior quantities of interest, which include the posterior mean, posterior median, and 95% credible intervals. The same number of observations (20), which is randomly selected for each simulation, is set for the test set to get comparable prediction results.

3.4.3.4 Data locations

The simulated rectangle spatial domain is 10 by 5, and the simulated ranges for the random fields ($\mu_1(\mathbf{s})$, $\mu_2(\mathbf{s})$ and $\mu_3(\mathbf{s})$) are 4, 3 and 2, respectively, which is a comparable scale. But this might not be the case when we use the real data. When working with spatial data, one common issue is when the spatial domain and the range of the spatial process are not on a comparable scale, which will lead to numerical problems and difficulties in obtaining accurate and stable results with INLA.

Figure 3.4 shows one replicate simulated location for y_1, y_2 and y_3 respectively. y_1, which is the misaligned covariate, its locations are represented by the red circles. y_2 and y_3, which are the non-misaligned covariate and the response variable, respectively, their locations are represented by the blue squares. The locations of the response variable in the test set are represented by the green triangles.

Since one of the big challenges of this project is that the monitoring network is sparse, the soil moisture sensors are located around the river instead of being evenly distributed in the study catchment. One scenario is to replicate the real sensor locations as the simulated locations to resemble what actual data looks like.

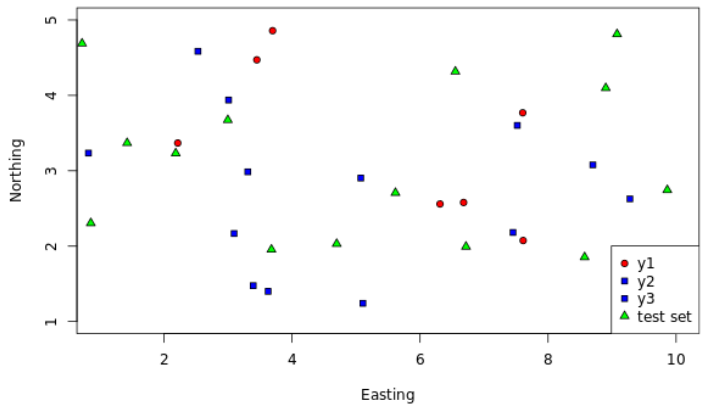


Figure 3.4: Simulated locations for the response variable and non-misaligned covariate (blue squares), misaligned covariate (red circles), and response variable in the test set (green triangles).

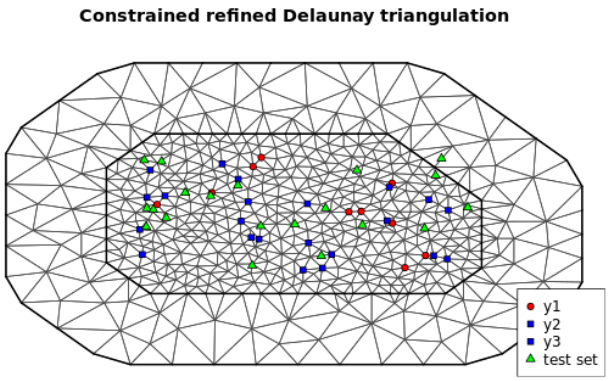


Figure 3.5: Mesh for the misaligned data. Blue and red dots denote the response and covariate locations, respectively. Green dots denote the test set of the response variable.

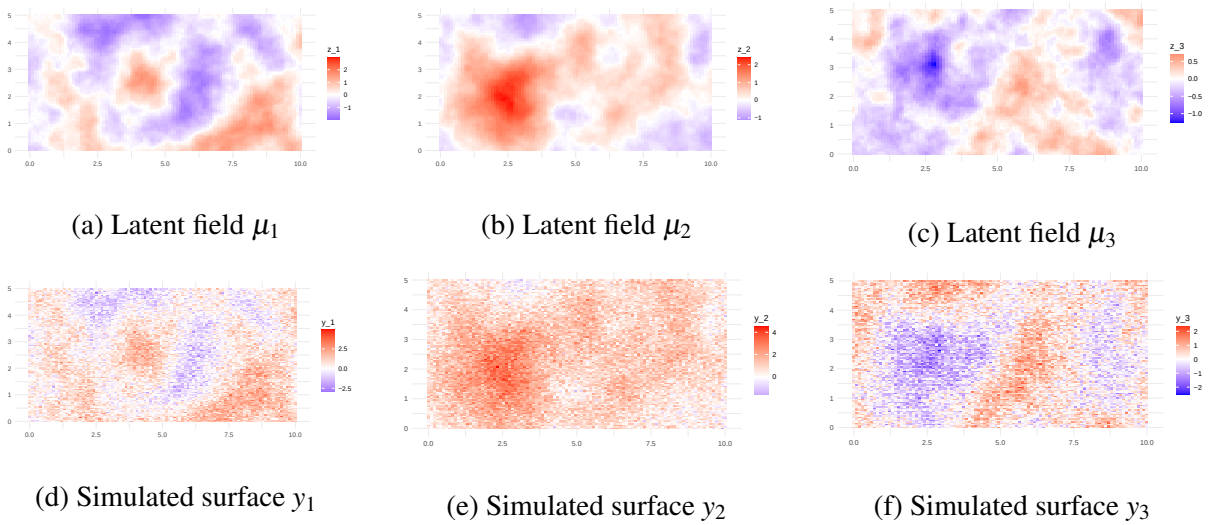
Figure 3.5 shows the mesh constructed from the simulated locations in Figure 3.4. The details of the mesh construction can be found in Section 3.3.3.4. The mesh is made of 478 points. The same mesh will be used to build the SPDE model for considering all three spatially structured random fields, which makes it easier to link different random effects across different outcomes at different spatial locations.

To be specific, the spatial domain refers to the extent of the area over which the spatial process is observed, while the range of the spatial process refers to the distance at which spatial correlation becomes negligible. If the spatial domain and the range are not on a comparable scale, it means that the spatial correlation pattern changes rapidly over the study area. When this happens, INLA may encounter problems during the approximation process, which could lead to issues such as slow convergence, large uncertainties in the estimates, and potential biases in the results. The nu-

merical challenges are mainly because the INLA methodology relies on Laplace approximations and sparse matrices to handle the computational burden of Bayesian spatial models.

3.4.3.5 Simulated surfaces

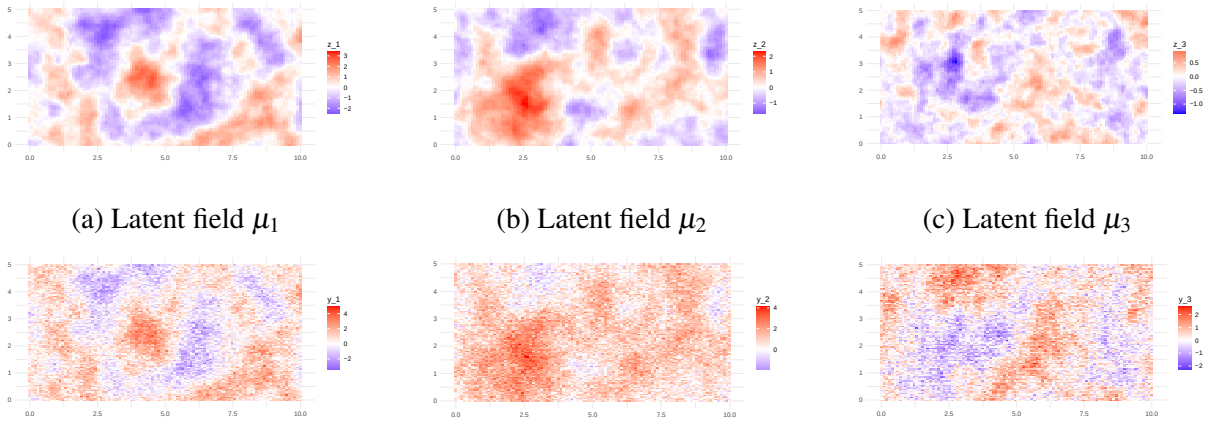
Figure 3.6, Figure 3.7, and Figure 3.8 present a single realization of the true exposure surfaces, for $\mu_1(\mathbf{s}), \mu_2(\mathbf{s}), \mu_3(\mathbf{s})$ respectively, and simulated surfaces, for $y_1(\mathbf{s}), y_2(\mathbf{s}), y_3(\mathbf{s})$ respectively, derived from each of the aforementioned scenarios.



(a) Latent field μ_1 (b) Latent field μ_2 (c) Latent field μ_3

(d) Simulated surface y_1 (e) Simulated surface y_2 (f) Simulated surface y_3

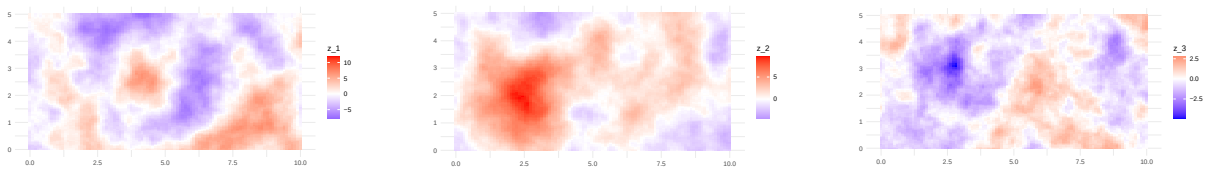
Figure 3.6: Simulated surfaces for latent field μ and simulated surface y in Scenario1



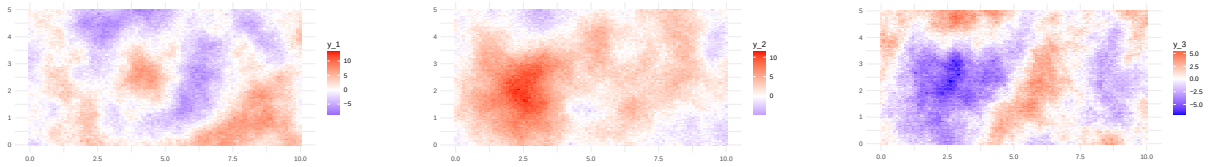
(a) Latent field μ_1 (b) Latent field μ_2 (c) Latent field μ_3

(d) Simulated surface y_1 (e) Simulated surface y_2 (f) Simulated surface y_3

Figure 3.7: Simulated surface for latent field μ and simulated surface y in Scenario2

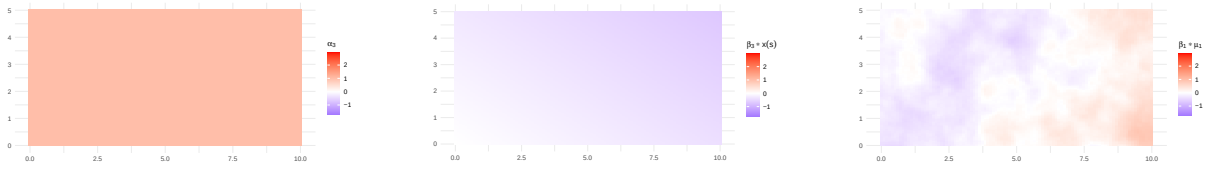


(a) Latent field μ_1 (b) Latent field μ_2 (c) Latent field μ_3

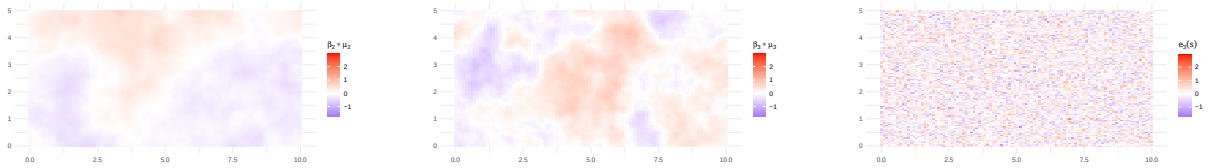


(d) Simulated surface y_1 (e) Simulated surface y_2 (f) Simulated surface y_3

Figure 3.8: Simulated surfaces for latent field μ and simulated surface y in Scenario3



(a) intercept α_3 (b) $\beta_3 * x(\mathbf{s})$ (c) $\beta_1 * (\alpha_1 + \mu_1)$



(d) $\beta_2 * (\alpha_2 + \mu_2)$ (e) μ_3 (f) $e_3(\mathbf{s})$

Figure 3.9: Simulated surfaces for latent field μ and simulated surface y in Scenario3

It is noted that in Scenario 1, Figure 3.6a and Figure 3.6d exhibit very similar patterns, as do Figure 3.6b and Figure 3.6e. However, Figure 3.6c and Figure 3.6f display notably different patterns. This difference arises from the composition of y_1 and y_2, which only include intercepts, latent fields, and measurement errors, while y_3 includes these elements along with fixed effects and shared random effects with y_1 and y_2. All components in Equation (3.6) contributing to y_3 are shown in Figure 3.9.

3.4.3.6 Model fitting

Table 3.3 shows the priors used in the joint model. The SPDE model will consider the PC-priors for the model parameters in the range $\rho = \sqrt{0.8\nu}/\kappa$, and the marginal standard deviation.

Table 3.3: Priors specification for joint model parameters.

Parameters	Informative prior	Non-informative prior
α_1		N(0,10)
α_2		N(0,10)
α_3		N(0,10)
β_1		N(0,10)
β_2		N(0,10)
β_3		N(0,10)
ρ_1	$\text{PC}(\rho_0, \alpha)$	
ρ_2	$\text{PC}(\rho_0, \alpha)$	
ρ_3	$\text{PC}(\rho_0, \alpha)$	
σ_1^2	$\text{PC}(\sigma_0, \alpha)$	
σ_2^2	$\text{PC}(\sigma_0, \alpha)$	
σ_3^2	$\text{PC}(\sigma_0, \alpha)$	
$\sigma_{e_1}^2$		$\text{Inv Gamma}(1, 5e-5)$
$\sigma_{e_2}^2$		$\text{Inv Gamma}(1, 5e-5)$
$\sigma_{e_3}^2$		$\text{Inv Gamma}(1, 5e-5)$

Non-informative priors are used for all parameters, but more realistic informative priors, for example, to generate the posterior in a logical range for the parameters, can be used if the posterior mean of some parameters is far away from the true values.
InvGamma prior with a shape of 1 and inverse-scale of 0.00005, which is the default non-informative prior for the precision because the gamma distribution can be used as a conjugate prior for precision, is chosen for θ of the Gaussian error distribution, which can be parameterised to:

$$\begin{aligned} \text{mean} &= \frac{\text{a}}{\text{b}} \\ \text{variance} &= \frac{\text{a}}{\text{b}^2} \\ \theta &= \log(\text{Gamma}(\text{a}, \text{b})) \end{aligned}$$

The priors for the fixed effects (intercept and slope) and the scaling coefficients are Normal distributions with a mean of 0 and precision (0.001), which are the default priors with a large variance to ensure the prior provides minimal information.

The GRF with the Matérn covariance function in the model provides a ridge in the likelihood for the spatial parameters range and marginal variance, which might cause overfitting by estimating spurious spatial trends or spurious temporal trends. The penalised-complexity (PC) priors are used here for the scale parameter (σ^2) and the spatial variances (ρ) of the Matérn GRFs, with the prior median marginal variance $P(\sigma > \sigma_0) = 0.05$ and the prior median range $P(\rho > \rho_0) = 0.5$, respectively. The penalised-complexity (PC) priors penalise complexity and the distance from the base model by shrinking the range toward infinity and the marginal variance toward zero (Fuglstad et al., 2019).

The mean of the standard deviations of y_1, y_2 and y_3, and the mean of ρ_1, ρ_2 and ρ_3 is used as the upper and lower limit of σ^2 and range individually, and the tail probability $\alpha = 0.5$.

3.4.3.7 Data structure

In the joint model, the response $\mathbf{y}_3(\mathbf{s})$, varies over space and depends on a misaligned covariate $\mathbf{y}_1(\mathbf{s})$ and an aligned covariate $\mathbf{y}_2(\mathbf{s})$, which also vary spatially. The covariates $\mathbf{y}_1(\mathbf{s})$ and $\mathbf{y}_2(\mathbf{s})$ are assumed to be observed with error.
Copying part of a linear predictor is needed for all joint modelling. The key point is the need to compute:

$$
\begin{aligned}
0 &= \eta_1(\mathbf{s}) + e_1 - \mathbf{y}_1(\mathbf{s}) \\
0 &= \eta_2(\mathbf{s}) + e_2 - \mathbf{y}_2(\mathbf{s}) \\
0 &= \alpha_3 + \beta_3 \mathbf{x}(\mathbf{s}) + \eta_1(\mathbf{s}) + \eta_2(\mathbf{s}) + e_3 - \mathbf{y}_3(\mathbf{s})
\end{aligned}
$$

from the first and second observation equations, to copy them to the third observation equations. So, a model that computes $\eta_1(\mathbf{s})$ and $\eta_2(\mathbf{s})$ explicitly needs to be defined.

The way we choose is to minimise the size of the graph generated by the model (Rue et al. 2017). First, the following equations are considered:

$$
\begin{aligned}
\mathbf{0}(\mathbf{s}) &= \eta_1(\mathbf{s}) + e_1 - \mathbf{y}_1(\mathbf{s}) \\
\mathbf{0}(\mathbf{s}) &= \eta_2(\mathbf{s}) + e_2 - \mathbf{y}_2(\mathbf{s}) \\
\mathbf{0}(\mathbf{s}) &= \alpha_3 + \beta_3 \mathbf{x}(\mathbf{s}) + \eta_1(\mathbf{s}) + \eta_2(\mathbf{s}) + e_3 - \mathbf{y}_3(\mathbf{s})
\end{aligned}
$$

where only $\mathbf{y}_1(\mathbf{s})$, $\mathbf{y}_2(\mathbf{s})$ and $\mathbf{y}_3(\mathbf{s})$ are known. For the $\eta_1(\mathbf{s})$ and $\eta_2(\mathbf{s})$ terms, we assume independent and identically distributed models with low fixed precision. With this fixed high variance, each element in $\eta_1(\mathbf{s})$ and $\eta_2(\mathbf{s})$ can take any value. However, these values will be forced to be $\alpha_1 + \mu_{\mathbf{1}}(\mathbf{s})$ and $\alpha_2 + \mu_{\mathbf{2}}(\mathbf{s})$ by considering a Gaussian likelihood for the "faked zero" observations with a high fixed precision value.

The joint model is described with five likelihoods. The data block matrix $\mathbf{D}$ is defined as follows, with the number of columns corresponding to the number of likelihoods and each block corresponding to the data used to estimate one of the linear predictors.

$$\mathbf{D} = \begin{pmatrix} \text{NA} & \begin{pmatrix} y_1(1) \\ \vdots \\ y_1(s_1) \end{pmatrix} & \text{NA} & \text{NA} & \text{NA} \\ \text{NA} & \text{NA} & \text{NA} & \begin{pmatrix} y_2(1) \\ \vdots \\ y_2(s_2) \end{pmatrix} & \text{NA} \\ 0 & \text{NA} & 0 & \text{NA} & \begin{pmatrix} y_3(1) \\ \vdots \\ y_3(n) \\ y_{3v}(n+1) \\ \vdots \\ y_{3v}(n+m) \end{pmatrix} \end{pmatrix}$$

The dimension of $\mathbf{D}$ is $(n_1+n_2+n_3+m)*5$, with $s_1 = n_1$, $s_2 = n_2$, $s_3 = n_3$, $m = 20$. The projection matrix is also associated with the test set. The prediction is computed at each unknown location, and the value of m depends on what prediction is being computed (either a test site or a regular grid in the fine-resolution prediction map).

3.4.3.8 Evaluation Metric

The performance of the joint model with misaligned covariate is evaluated by looking at the root mean squared error (RMSE) of the parameters, Equation (3.8), and the root mean squared prediction error (RMSPE) of the test set, Equation (3.9). The difference between RMSE and RMSPE is that RMSE involves only training data, while RMSPE involves a testing set that was not part of the model training. The RMSE_θ and RMSPE_y are computed as follows:

$$\text{RMSE}_\theta = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\theta_i - \hat{\theta}_i)^2}, \quad n = 1, ..., 100 \tag{3.8}$$

$$\text{RMSPE}_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \quad n = 1, ..., 20 \tag{3.9}$$

3.4.3.9 Results for spatio-only model

For each of the simulated surfaces, the posterior means and 95% credible intervals (CIs) for the parameters are calculated in Table 3.4, Table 3.5, and Table 3.6. The method doesn't accurately recover all the true values used in the simulation, but the 95% CIs contain the true value for all parameters.

Table 3.4: Mean of the posterior distributions of the parameters in the spatio-only model (scenario1)

	True	Mean	0.975quant	0.025quant
α_1	0.5	0.59	1.09	0.09
α_2	0.8	0.87	1.14	0.59
α_3	1	0.59	1.3	-0.12
β_1	-0.3	-0.12	0.28	-0.53
β_2	-0.4	-0.08	0.21	-0.38
β_3	-0.2	-0.19	0.18	-0.57
$\sigma^2_{e_1}$	0.36	0	0	0
$\sigma^2_{e_2}$	0.25	0	0.01	0
$\sigma^2_{e_3}$	0.16	0	0.02	0
ρ_1	4	2.25	4.64	0.9
ρ_2	3	1.34	2.46	0.66
ρ_3	2	1.93	5.35	0.63
σ_1	1	0.86	1.26	0.58
σ_2	0.5	0.64	0.87	0.47
σ_3	0.3	0.46	0.7	0.3

Table 3.5: Mean of the posterior distributions of the parameters in the spatial-only model (scenario2)

	True	Mean	0.975quant	0.025quant
α_1	0.5	0.42	0.97	-0.12
α_2	0.8	0.78	1.05	0.52
α_3	1	0.67	1.44	-0.1
β_1	-0.3	-0.15	0.21	-0.52
β_2	-0.4	-0.09	0.21	-0.38
β_3	-0.2	-0.18	0.23	-0.6
$\sigma^2_{e_1}$	0.36	0	0	0
$\sigma^2_{e_2}$	0.25	0	0	0
$\sigma^2_{e_3}$	0.16	0	0.02	0
ρ_1	2	1.8	3.63	0.74
ρ_2	1.5	1.17	2.18	0.54
ρ_3	1	3.48	14.87	0.67
σ_1	1	0.97	1.39	0.67
σ_2	0.5	0.65	0.88	0.47
σ_3	0.3	0.44	0.7	0.27

Table 3.6: Mean of the posterior distributions of the parameters in the spatial-only model (scenario3)

	True	Mean	0.975quant	0.025quant
α_1	0.5	0.71	1.95	-0.53
α_2	0.8	1.07	1.7	0.45
α_3	1	0.88	2.53	-0.77
β_1	-0.3	-0.33	-0.08	-0.58
β_2	-0.4	-0.23	0.04	-0.5
β_3	-0.2	-0.19	0.68	-1.06
$\sigma^2_{e_1}$	0.36	0	0	0
$\sigma^2_{e_2}$	0.25	0	0	0
$\sigma^2_{e_3}$	0.16	0	0	0
ρ_1	4	1.89	3.18	1.04
ρ_2	3	1.96	3.27	1.09
ρ_3	2	4.71	19.92	0.86
σ_1	4	2.26	2.96	1.71
σ_2	2	1.48	1.95	1.12
σ_3	1.2	0.52	1.07	0.26

Table 3.4, 3.5, and 3.6 display the mean values from the posterior distributions of parameters within the spatial-only model for each scenario. Notably, the precision parameter consistently approaches 0, which means the measurement error is close to 0 due to the limited data size. It's worth highlighting that the scaling parameters β_1 and β_2 exhibit better estimations in Scenario 3 compared to Scenarios 1 and 2. This improvement arises from the larger marginal standard deviation in scenario 3, resulting in a more varied latent field. Consequently, evaluating the scaling parameter between the two latent fields becomes comparatively easier in this scenario.

Table 3.7: RMSE_θ of all parameters in the spatio-only model for scenarios1,2,3

	Scenario1	Scenario2	Scenario3
α_1	2.48	2.61	3.18
α_2	4.14	4.22	4.01
α_3	3.45	3.36	3.51
β_1	0.28	0.29	0.35
β_2	0.35	0.35	0.22
β_3	0.21	0.21	0.74
$\sigma^2_{e_1}$	0.36	0.36	0.36
$\sigma^2_{e_2}$	0.25	0.25	0.25
$\sigma^2_{e_3}$	0.16	0.16	0.16
ρ_1	2.05	0.91	2.27
ρ_2	1.76	0.55	1.23
ρ_3	1.63	12.66	5.84
σ_1	0.23	0.2	1.81
σ_2	0.18	0.18	0.58
σ_3	0.2	0.22	0.86

Table 3.7 shows the RMSE_{θ}s of all parameters within the spatial-only model calculated from Equation (3.8). It suggests that the accuracy of parameter estimation is influenced by both the characteristics of the simulated surfaces and the number of locations involved. In the real data application, the spatial parameters that control the degree of smoothing of the surface are unknown. The bias occurs in scenarios that have sparse monitoring data or the variance is correlated with confounders and response variables at small spatial scales (Gryparis et al., 2009). The accuracy of parameter estimation is influenced by both the characteristics of the simulated surfaces and the number of locations involved.

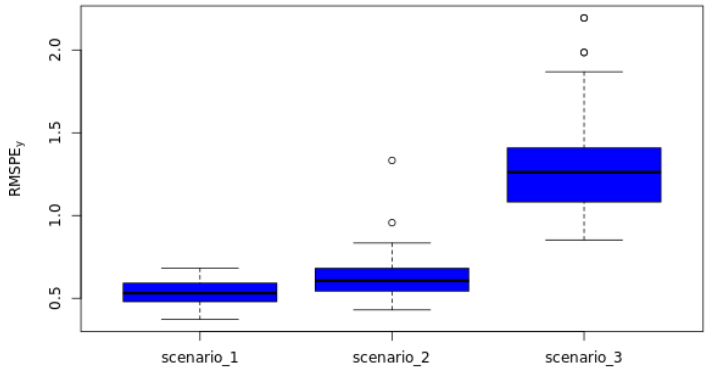


Figure 3.10: Performance comparison between models in each scenario

Figure 3.10 demonstrates the changes in RMSPE_y calculated from Equation (3.9) for the test set across different scenarios. The predictive outcomes indicate better performance when dealing with a more flattened latent field. In Scenario 2, the latent field has a smaller range compared to Scenario 1, while Scenario 3 has a latent field with a greater marginal deviation than Scenario 1. Both scenarios with less flattened latent fields show higher RMSPE_y values than Scenario 1, indicating worse predictive results.

Table 3.8: Mean of the posterior distributions of the parameters in the spatio-only model (scenarios with varying number of locations)

	True / Sample size	10_22_22	20_44_44	40_88_88	80_176_176	100_220_220	200_440_440
α_1	0.5	0.46	0.44	0.49	0.49	0.47	0.46
α_2	0.8	0.76	0.79	0.8	0.79	0.78	0.79
α_3	1	0.76	0.68	0.74	0.82	0.84	0.94
β_1	-0.3	-0.15	-0.18	-0.22	-0.24	-0.24	-0.3
β_2	-0.4	-0.09	-0.12	-0.15	-0.24	-0.31	-0.35
β_3	-0.2	-0.25	-0.19	-0.19	-0.2	-0.17	-0.19
$\sigma^2_{e_1}$	0.36	0	0	0.02	0.18	0.25	0.34
$\sigma^2_{e_2}$	0.25	0.01	0.02	0.08	0.22	0.24	0.24
$\sigma^2_{e_3}$	0.16	0.01	0.02	0.07	0.13	0.14	0.16
ρ_1	4	2.19	1.93	1.68	2.57	3.43	3.58
ρ_2	3	1.37	1.41	1.68	3.33	3.18	3.38
ρ_3	2	2.15	2.44	3.51	2.78	2.74	2.7
σ_1	1	0.88	0.95	0.99	0.95	0.92	0.88
σ_2	0.5	0.63	0.67	0.61	0.5	0.47	0.5
σ_3	0.3	0.47	0.45	0.39	0.33	0.31	0.3

Table 3.9: Empirical coverage of 95% credible intervals for all parameters in the spatial-only model (scenarios with a varying number of locations).

Parameter / Sample size	10_22_22	20_44_44	40_88_88	80_176_176	100_220_220	200_440_440
α_1	0.61	0.60	0.50	0.24	0.25	0.18
α_2	0.89	0.67	0.54	0.38	0.42	0.25
α_3	0.74	0.75	0.70	0.74	0.69	0.81
β_1	0.80	0.87	0.75	0.84	0.87	0.85
β_2	0.41	0.56	0.32	0.75	0.89	0.91
β_3	0.87	0.86	0.87	0.90	0.91	0.92
$\sigma^2_{e_1}$	0.00	0.32	0.10	0.54	0.75	0.96
$\sigma^2_{e_2}$	0.04	0.11	0.39	0.97	0.92	0.93
$\sigma^2_{e_3}$	0.02	0.23	0.47	0.87	0.89	0.95
ρ_1	0.43	0.34	0.16	0.49	0.80	0.87
ρ_2	0.33	0.34	0.35	0.81	0.91	0.98
ρ_3	0.76	0.84	0.62	0.91	0.92	0.92
σ_1	0.89	0.82	0.85	0.90	0.88	0.82
σ_2	0.57	0.55	0.47	0.93	0.89	0.95
σ_3	0.35	0.42	0.51	0.87	0.93	0.97

Table 3.8 presents the mean from the posterior distributions of the parameters within the spatial-only model. It reveals consistent and accurate estimations for the fixed effects $(\alpha_1, \alpha_2, \alpha_3, \beta_3)$ regardless of the amount of available data. However, for the remaining parameters, for example,

$\sigma^2_{e_1}$, $\sigma^2_{e_2}$, $\sigma^2_{e_3}$ are close to zero in scenarios 1, 2, and 3, which suggests that insufficient data from a restricted number of locations leads to failure of accurately recovering the true values.

The RMSE$_\theta$s for range parameters ρ are considerably large compared with other parameters within the model, but the posterior mean is acceptable, this discrepancy between the mean and RMSE$_\theta$ could caused by outliers, despite being balanced out in the mean, contributing significantly to the overall squared error when calculating the RMSE$_\theta$. While the average estimate is near the truth, substantial individual errors might increase the overall RMSE$_\theta$.

Table 3.9 shows the empirical coverage of the 95% credible intervals. For many parameters, especially the β coefficients and several variance terms, the coverage is close to the nominal 0.95 level as the number of locations increases, which is consistent with the decreasing RMSE and narrower intervals. In contrast, the coverage for α_1 and α_2 is clearly below 0.95 in the larger-sample scenarios, suggesting some remaining bias in these components. The range parameters also show poorer coverage in smaller samples, which reflects that they are more difficult to identify.

Table 3.10: Root mean squared error (RMSE$_\theta$) of parameter estimates in the spatial-only model (scenarios with a varying number of locations).

Parameter \ Sample size	10_22_22	20_44_44	40_88_88	80_176_176	100_220_220	200_440_440
α_1	0.56	0.57	0.46	0.47	0.53	0.46
α_2	0.26	0.21	0.22	0.23	0.20	0.19
α_3	0.56	0.53	0.32	0.33	0.33	0.24
β_1	0.25	0.21	0.17	0.13	0.12	0.08
β_2	0.35	0.32	0.29	0.23	0.2	0.14
β_3	0.25	0.2	0.15	0.13	0.14	0.1
$\sigma^2_{e_1}$	0.36	0.36	0.35	0.25	0.17	0.05
$\sigma^2_{e_2}$	0.25	0.23	0.2	0.05	0.04	0.02
$\sigma^2_{e_3}$	0.16	0.15	0.11	0.05	0.04	0.01
ρ_1	2.15	2.28	2.5	2.09	1.83	1
ρ_2	1.73	2.29	2.04	1.58	1.41	1.09
ρ_3	2.3	3.63	8.42	1.64	1.65	1.9
σ_1	0.22	0.18	0.17	0.15	0.16	0.19
σ_2	0.17	0.2	0.15	0.09	0.09	0.08
σ_3	0.22	0.19	0.17	0.1	0.08	0.05

Table 3.11: Mean 95% credible interval width for all parameters in the spatial-only model (scenarios with a varying number of locations).

Parameter \ Sample size	10_22_22	20_44_44	40_88_88	80_176_176	100_220_220	200_440_440
α_1	1.04	0.87	0.60	0.43	0.39	0.25
α_2	0.56	0.44	0.29	0.24	0.20	0.13
α_3	1.43	1.07	0.91	0.82	0.77	0.69
β_1	0.79	0.67	0.43	0.36	0.37	0.32
β_2	0.58	0.49	0.46	0.59	0.59	0.49
β_3	0.78	0.65	0.50	0.46	0.43	0.38
$\sigma^2_{e_1}$	0.54	0.42	0.24	0.21	0.26	0.20
$\sigma^2_{e_2}$	0.02	0.05	0.13	0.07	0.12	0.08
$\sigma^2_{e_3}$	0.01	0.07	0.09	0.06	0.09	0.06
ρ_1	3.15	3.31	1.80	3.10	4.40	4.02
ρ_2	2.13	3.68	2.50	4.98	4.57	4.17
ρ_3	6.17	8.64	12.78	7.31	5.64	4.67
σ_1	0.66	0.62	0.51	0.53	0.60	0.58
σ_2	0.41	0.40	0.30	0.41	0.34	0.34
σ_3	0.39	0.32	0.31	0.30	0.27	0.23

Table 3.10 shows all the RMSE$_\theta$s calculated from Equation (3.8) across scenarios involving different numbers of locations (scenarios 1, 4, 5, 6, 7, 8). Scenario 8 consistently has the smallest RMSE$_\theta$, indicating that incorporating more data can reduce bias in parameter estimation. While the RMSE$_\theta$s for the intercepts $(\alpha_1, \alpha_2, \alpha_3)$ demonstrate marginal differences, those for other parameters display considerable variations.

Table 3.11 reports the mean 95% credible interval width for all parameters in the spatial-only model as the number of locations increases. For most regression coefficients and variance parameters, the intervals become narrower with a larger number of sensors, indicating that the posterior distributions are more concentrated and the parameters are estimated more precisely when more spatial information is available. In contrast, some parameters, especially the range parameters, still have relatively wide intervals or unstable behaviour across scenarios, which suggests that these components are harder to identify from the data and are more sensitive to the modelling assumptions.

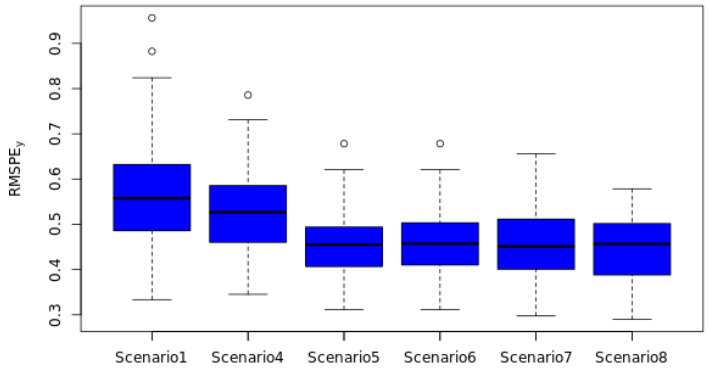


Figure 3.11: Performance comparison between models in each scenario

Figure 3.11 illustrates how the RMSPE$_y$ values calculated from Equation (3.9) for the test set decrease as the number of points increases. This trend is logical since the number of locations helps with enhancing the parameter estimation of the Matérn field, thereby enhancing predictive accuracy. To be more specific, the trend experiences a significant decline as the number of locations rises from 10_22_22 to 80_176_176, after which it stabilises at that level. It is noted that there are two outliers in scenario 1 and one outlier each appears in scenarios 2 and 4, which might suggest more replicates are needed for the simulation. Overall, looking at RMSE, interval width and coverage together gives a more complete picture of how well the spatial-only model recovers the true parameters and helps motivate the more flexible spatio-temporal data-fusion models in the following chapters.

3.4.3.10 Conclusions

The simulation study of the spatial-only model suggests that, for each of the simulated surfaces, the posterior means and 95% CIs for the parameters are calculated in Table 3.4, Table 3.5, and Table 3.6, and Table 3.8. The method only accurately recovers the true values of β_1, $\sigma^2_{e_3}$, σ_2, and σ_3, but the 95% CIs contain the true value for most of the parameters. Figure 3.11 shows that predictive performance and the parameter estimation exhibit the expected improvement with an increasing number of locations. The trend experiences a significant decline as the number of locations rises from 10_22_22 to 80_176_176, after which it stabilises at that level. Section 3.4.4 presents the simulation of the spatio-temporal model indicated by Equation (3.7), aiming to understand the performance of parameter estimation within the model.

3.4.4 Spatio-temporal model

The spatio-temporal model is defined in Equation (3.7). The simulation study of the spatio-temporal model has the same dependency structure as the spatio-only model. It extends the

spatio-only model by computing the latent field at each time point conditionally on the previous one.

3.4.4.1 Parameters for data simulation

The parameters for data simulation in the spatio-temporal model are the same as the setting of the spatio-only model, while for the spatio-temporal dependency, it introduces the temporal coefficients a_1, a_2 and a_3. All the other parameters utilise identical settings within the spatial-only Scenario 3.

3.4.4.2 Simulation strategy

The simulation of the spatio-temporal model is an extension of the spatial-only model simulation. The first four steps are the same as the simulation of the spatial-only model. Step 5 is to generate multiple latent fields for each time point.

5. The temporal correlation is introduced by the formula as follows:

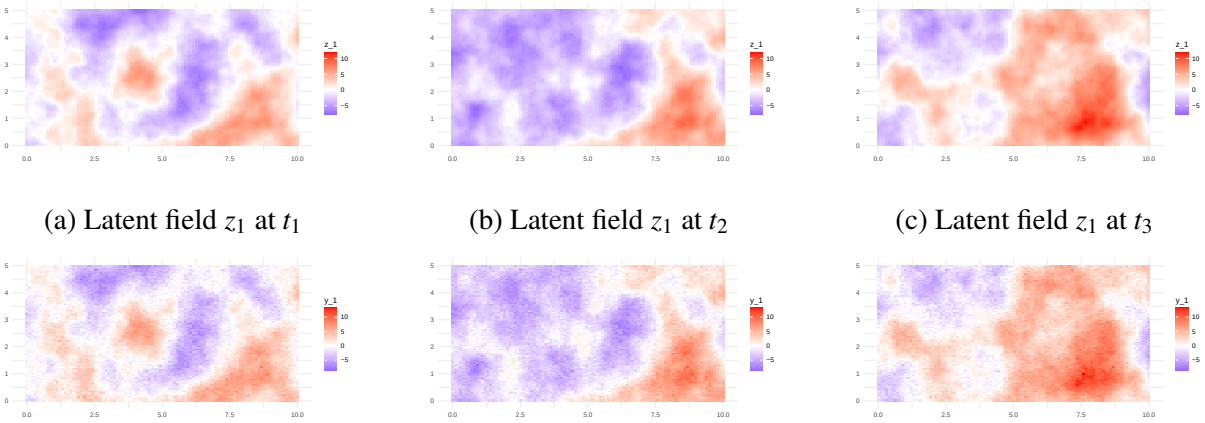
$$\eta(s,t) = a * \eta(s,t-1) + \sqrt{(1-a^2)} * \eta(s,t),$$

   where $\sqrt{1-a^2}$ term is used to make the process stationary in time. The spatio- temporal process is assumed to be a series of GRFs, and the latent process accounts for the temporal dependencies using the AR(1) model.

6. Design for the test set: The testing approach for evaluating the spatio-temporal model can be outlined in two ways:
   Firstly, for every scenario within the spatio-temporal context, 20 random locations are chosen daily. The Root Mean Squared Prediction Error (RMSPE) is then computed to assess the model's prediction performance.
   Alternatively, another method involves selecting 20 points on day k+1 and computing the RMSPE for these specific points on the subsequent unseen day.
   The final aim of this study is to create a high-resolution map with the limited data available. Therefore, we chose the first test set design.

Spatio-temporal data visualisation

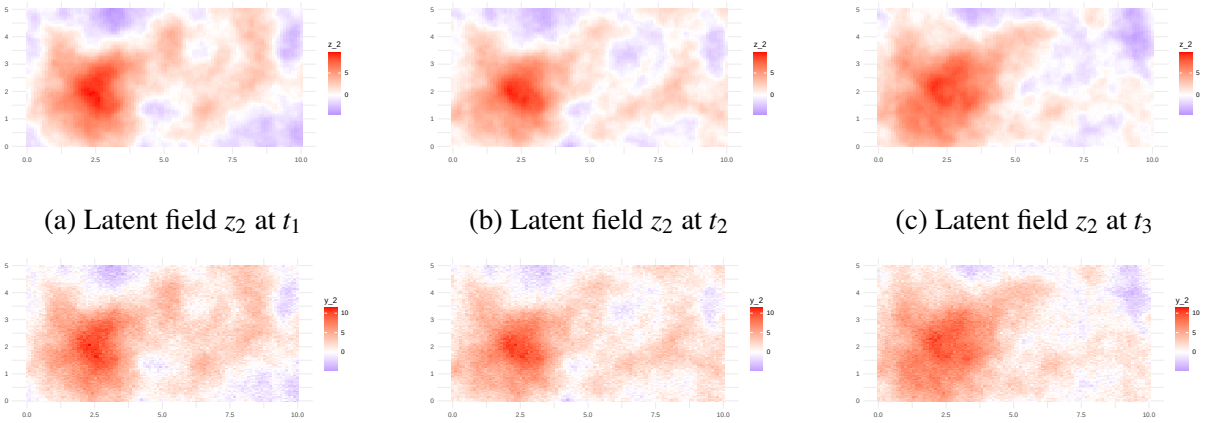
Figure 3.12, Figure 3.13, and Figure 3.14 show the realisation of the space-time process in 3 time points for each variable. The spatio-temporal process is constructed by generating t-independent realisations of the spatial model, where t is the number of time points. The temporal process in the simulation study is assumed to be the AR(1) process.



(a) Latent field z_1 at t_1 (b) Latent field z_1 at t_2 (c) Latent field z_1 at t_3

(d) Simulated surface y_1 at t_1 (e) Simulated surface y_1 at t_2 (f) Simulated surface y_1 at t_3

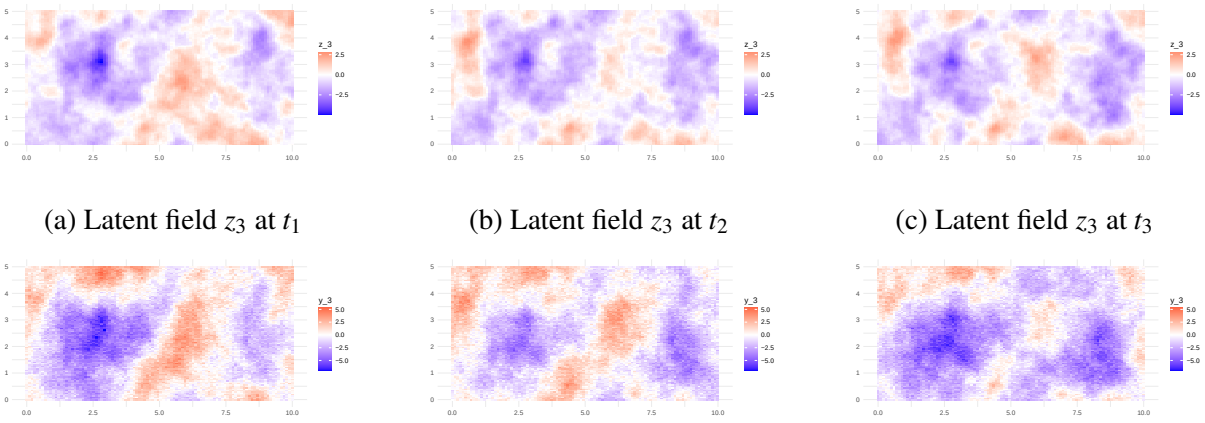
Figure 3.12: Realisation of the space-time random field and y_1



(a) Latent field z_2 at t_1 (b) Latent field z_2 at t_2 (c) Latent field z_2 at t_3

(d) Simulated surface y_2 at t_1 (e) Simulated surface y_2 at t_2 (f) Simulated surface y_2 at t_3

Figure 3.13: Realisation of the space-time random field and y_2



(a) Latent field z_3 at t_1 (b) Latent field z_3 at t_2 (c) Latent field z_3 at t_3

(d) Simulated surface y_3 at t_1 (e) Simulated surface y_3 at t_2 (f) Simulated surface y_3 at t_3

Figure 3.14: Realization of the space-time random field and y_3

This series of figures provides a comprehensive visualisation of spatial-temporal dynamics using a Matern covariance function for spatial variation and varying temporal coefficients to model change over time.

Figure 3.12 displays a relatively rapid rate of change in the spatial field over time with a temporal coefficient of 0.7. Each successive figure shows a more noticeable change from the initial state, suggesting more significant changes over time. Figure 3.13 displays moderate temporal dynamics, this figure reveals how the spatial field evolves with a temporal coefficient of 0.8, resulting in a slower rate of change compared to Figure 3.12. Figure 3.14 displays slow temporal dynamics, revealing how the spatial field evolves with a temporal coefficient of 0.9. The evolution is gradual, with each figure showing subtle noticeable changes from the initial state. This high rate of change allows for an understanding of the evolving spatial patterns without drastic transformations.

3.4.4.3 Results for spatio-temporal model

Table 3.12: Mean of the posterior distributions of the parameters in the spatio-temporal model with 95% CIs

	True	k=1	k=3	k=15	k=30
α_1	0.5	0.57(-0.12,1.27)	0.57(0.04,1.09)	0.51(0.19,0.82)	0.53(0.28,0.78)
α_2	0.8	0.74(0.39,1.09)	0.81(0.56,1.06)	0.79(0.61,0.98)	0.81(0.68,0.94)
α_3	1	0.64(-0.14,1.42)	0.76(0.23,1.29)	0.92(0.63,1.21)	1.02(0.82,1.23)
β_1	-0.3	-0.16(-0.58,0.24)	-0.24(-0.46,-0.02)	-0.27(-0.38,-0.16)	-0.3(-0.38,-0.23)
β_2	-0.4	-0.11(-0.4,0.18)	-0.15(-0.36,0.06)	-0.32(-0.49,-0.14)	-0.39(-0.51,-0.28)
β_3	-0.2	-0.17(-0.57,0.23)	-0.21(-0.49,0.06)	-0.2(-0.35,-0.05)	-0.22(-0.33,-0.11)
$\sigma^2_{e_1}$	0.36	0(0,0)	0(0,0.01)	0.08(0.05,0.16)	0.21(0.14,0.35)
$\sigma^2_{e_2}$	0.25	0.01(0.01,0.04)	0.03(0.02,0.08)	0.2(0.15,0.28)	0.21(0.18,0.26)
$\sigma^2_{e_3}$	0.16	0(0,0)	0.04(0.02,0.09)	0.1(0.07,0.15)	0.12(0.09,0.16)
ρ_1	4	1.82(0.78,3.69)	2.25(1.21,3.83)	2.85(2.06,3.87)	3.38(2.57,4.34)
ρ_2	3	1.36(0.65,2.56)	1.57(0.81,2.9)	3.21(2.05,4.86)	3.09(2.4,3.98)
ρ_3	2	2.04(0.7,5.8)	2.63(0.86,6.76)	2.37(1.06,4.85)	3.3(1.08,9.71)
σ_1	1	0.91(0.62,1.32)	1(0.76,1.28)	1.05(0.91,1.21)	1(0.89,1.12)
σ_2	0.5	0.63(0.46,0.86)	0.64(0.52,0.79)	0.52(0.43,0.62)	0.51(0.45,0.57)
σ_3	0.3	0.44(0.29,0.67)	0.41(0.27,0.6)	0.34(0.25,0.46)	0.31(0.23,0.42)
a_1	0.4	-	0.17(-0.15,0.48)	0.25(0.09,0.4)	0.32(0.19,0.43)
a_2	0.5	-	0.21(-0.08,0.48)	0.44(0.24,0.61)	0.46(0.34,0.57)
a_3	0.6	-	0.19(-0.28,0.6)	0.43(0.14,0.64)	0.5(0.29,0.65)

Table 3.12 displays the posterior mean and 95% credible intervals (CIs) for all parameters within the spatio-temporal model. In the scenario of a spatio-temporal model involving 3 time points, $\sigma^2_{e_1}$, $\sigma^2_{e_2}$, $\sigma^2_{e_3}$ are close to zero due to limited data availability, both spatially and temporally. The true values of fixed effects α_1, α_2, α_3, and β_3 all fall within the 95% CIs of their respective posterior distributions.

Regarding the scaling parameter β_1, it lies within the 95% CIs, whereas β_2 marginally extends toward the upper edge of this interval. As for the range parameters, ρ, ρ_1, and ρ_2 tend slightly toward the upper edge of the 95% CIs, while ρ_3 falls within the 95% CIs.

The marginal standard deviation σ exhibits variability, with σ_2 falling slightly below the lower limit of the 95% CIs, contrasted by σ_1 and σ_3 falling within this interval. As for the temporal coefficients, a_1 falls within the 95% CIs, whereas a_2 and a_3 closely approach the boundary. However, increasing the time points to 30 results in only a slight deviation of the true value of $\sigma^2_{e_1}$ from the 95% CIs, while other parameter estimations within the spatio-temporal model fall into the 95% CIs.

Table 3.13: RMSE$_\theta$ of all parameters in the spatio-temproal model for scenarios(a),(b),(c)

Parameters	k = 3	k = 15	k = 30
α_1	0.47	0.18	0.19
α_2	0.2	0.1	0.07
α_3	0.43	0.2	0.17
β_1	0.19	0.07	0.06
β_2	0.28	0.15	0.13
β_3	0.19	0.09	0.06
$\sigma^2_{e_1}$	2.06	1.68	1.2
$\sigma^2_{e_2}$	1.83	1.25	0.62
$\sigma^2_{e_3}$	2.67	1.83	5.3
ρ_1	0.36	0.32	0.24
ρ_2	0.23	0.1	0.07
ρ_3	0.13	0.09	0.07
σ_1	0.15	0.11	0.12
σ_2	0.17	0.07	0.05
σ_3	0.17	0.11	0.1
a_1	0.56	0.46	0.39
a_2	0.62	0.39	0.35
a_3	0.81	0.55	0.45

Table 3.12 displays the posterior mean of parameters in the spatio-temporal model, while Table 3.13 shows RMSE$_\theta$. The number of locations aligns with the sensor count for each variable in the real data, and there are 3, 15, and 30 time points for Scenarios a, b, and c, respectively. Comparing Scenario 1 in both the spatio-temporal and spatio-only models (Table 3.13 and Table 3.7), despite the spatio-temporal model having only 3 time points, the parameter estimation significantly improves. For instance, RMSE$_\theta$ of β_1 drops from 0.28 to 0.19, β_2 drops from 0.35 to 0.28, ρ_1 drops from 2.05 to 0.36, and α drops from 2.48 to 0.47. Nevertheless, precision error estimation remains notably far away from true values.

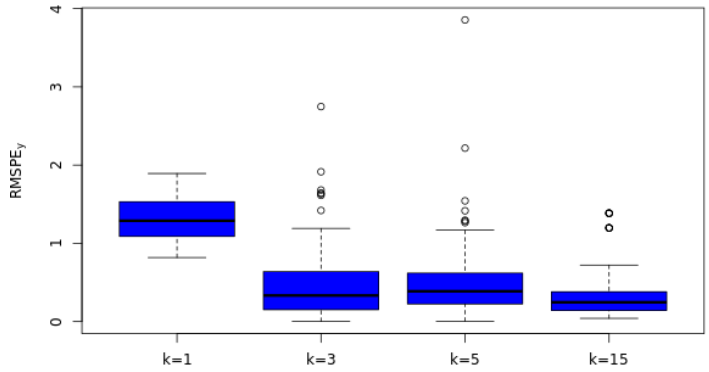


Figure 3.15: Performance comparison between models with different numbers of time points ($\rho_1 = 0.7, \rho_2 = 0.8, \rho_3 = 0.9$)

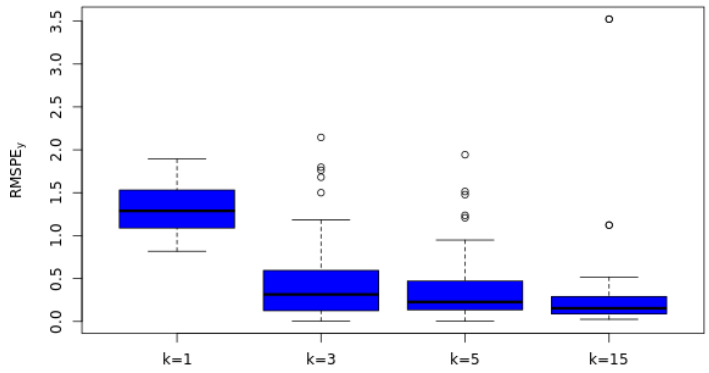


Figure 3.16: Performance comparison between models with different numbers of time points ($\rho_1 = 0.2, \rho_2 = 0.3, \rho_3 = 0.4$)

Figure 3.15 and Figure 3.16 illustrate how the RMSE$_\theta$ values calculated from Equation (3.9) for the test set decrease as the number of time points increases. This trend is logical since the number of time points helps with enhancing the parameter estimation of the Matérn field, thereby enhancing predictive accuracy. To be more specific, the trend experiences a significant decline as the number of time points increases from 1 to 15, and the variance of the RMSPE$_y$ decreases with the number of time points increasing. Comparing the case with k=1 in Figure 3.11 and the spatio-only case in Figure 3.16, it is noted that the RMSPE for the spatio-temporal model is not identical to that of the spatio-only model. Although the dependency structures of both models are the same, the spatio-temporal model's increased complexity needs more data to demonstrate proper performance.

3.5 Real data application

In this section, model (3.6) is applied to the SEPA data. The Coordinate Reference System (CRS) should be chosen based on the extent of the study area. It is important when working with multiple datasets that potentially have different coordinate systems, as they will need to be appropriately projected to the same CRS. For small areas, Easting-Northing coordinate systems are the most suitable choice. They effectively express the coordinates on a flat surface, which does not take into account the global curvature and consequent modification of the projection shape. The dataset in this study uses Easting/Northing coordinates and is projected using the local shape-preserving system British National Grid (BNG). VWC is standardised to 0 and 1 using the formula: $y = (x - \min(x))/(\max(x) - \min(x))$, and to satisfy the assumption of a Gaussian distribution, the logarithm transformation is applied to the response variable.

For the spatio-only model, the VWC sensor data, originally captured at a temporal resolution of every 15 minutes, are aggregated to a daily resolution to align them with the satellite data. The data used here corresponds to 06/05/2022, and the original spatial pattern and the prediction map are shown in Figure 3.17.

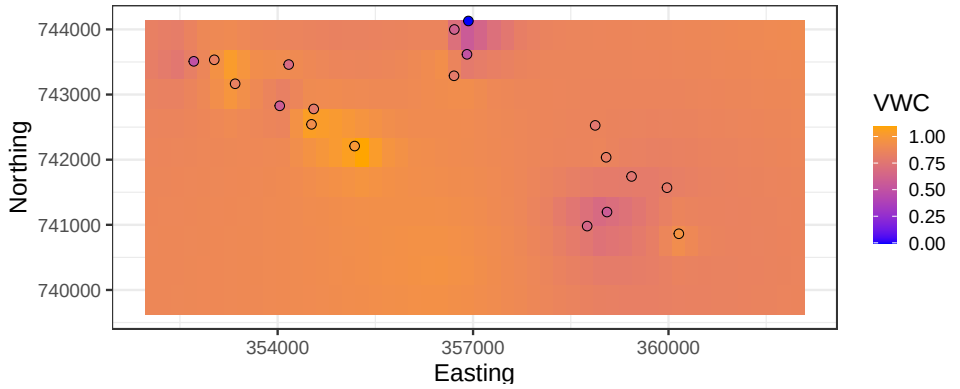


Figure 3.17: VWC from sensors in the Elliot water catchment on 06/05/2022

Model (3.6) is implemented for the soil moisture dataset of the Elliot Water catchment, where y_1 represents rainfall, y_2 represents soil temperature, y_3 represents Volumetric water content (VWC), and x denotes high resolution elevation data. Figure 3.17 displays the predicted soil moisture map for the Elliot water catchment. The circles represent the actual VWC values measured by sensors. Due to the sparse monitoring network of the in-situ sensors, the predicted mean does not exhibit significant spatial variation. The elevation, which is available everywhere, accounts for the observed spatial patterns in the areas where there are no sensors.

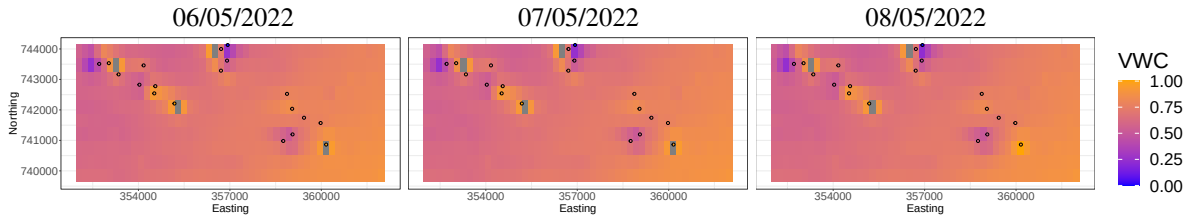


Figure 3.18: Prediction VWC from sensors in the Elliot Water catchment on 06/05/2022 (left), 07/05/2022 (middle) and 08/05/2022 (right)

Model (3.7) is implemented for the soil moisture dataset of the Elliot Water catchment, where y_1 represents rainfall, y_2 represents soil temperature, y_3 represents volumetric water content (VWC), and x denotes high-resolution elevation data. Figure 3.18 (right) displays the predicted soil moisture map for the Elliot water catchment. The circles represent the actual VWC values measured by sensors. Compared with the spatio-only prediction in Figure 3.17, model (3.7) gets a greater explanation of spatial variation by incorporating temporal data from the dataset.

3.6 Conclusions and discussion

For the spatio-only model (3.6), simulations calibrated to the real sensor network setting $(10, 22, 22)$ show that the 95% credible intervals (CIs) for the fixed effects achieve good coverage at this network density. By contrast, coverage for latent field hyperparameters (e.g., spatial range and marginal variance) is weaker. With relatively few sensors, the data are less informative about fine-scale spatial structure. As the sample size increases, CIs coverage improves across all parameters, which matches the behaviour seen in the large sample size scenario.

Compared to model (3.6), the spatio-temporal model (3.7) improves both parameter estimation and prediction. Increasing the time points from $k = 1$ to $k = 30$ obtains narrower CIs and reduced bias for the fixed effects and hyperparameters. Temporal replication at the same locations increases information about the latent process and helps estimate spatial variation and measurement noise.

In the real data application, despite these improvements, the predicted mean surfaces show limited fine-scale variability in regions far from sensors. This is expected with a sparse monitoring network: in unsampled areas, the model borrows strength from covariates. Elevation is available everywhere and explains most of the broad spatial variance, while local variations are weakly identified without nearby sensor support. This explains why uncertainty maps show wider intervals where elevation alone must carry the signal.

The contrast between $k = 1$ and $k = 30$ highlights a practical trade-off. Adding temporal points improves parameter estimates and improves predictions even without adding new sites, because

repeated observations at fixed locations constrain the latent field and noise components. However, temporal replication cannot fully compensate for spatial coverage: fine-scale spatial detail still requires measurements in space (more sites or complementary grids).

These results support using the spatio-temporal model (3.7) when more time points are available: it improves precision, calibration, and predictive performance over the spatial-only baseline (3.6). The residual lack of spatial richness away from sensors is a data coverage limitation rather than a modelling failure. A natural extension is to incorporate satellite gridded data into the model to increase spatial support between sensors, with the goal of sharpening spatial detail while maintaining calibrated uncertainty. We develop and evaluate the gridded and point fusion in Chapter 4.

Chapter 4

Data fusion method for the spatial only model

4.1 Introduction

This chapter presents a data fusion method to integrate high-resolution point data from SEPA in-situ sensors with lower-resolution grid data from Copernicus satellite images. The motivation for this method comes from the challenges of fusing soil moisture data collected from these two data sources from 2020 onwards. The integration of point data and grid data is an active area of research in remote sensing and spatial statistics, where combining data with different spatial resolutions enhances both accuracy and resolution by leveraging the trade-offs of each data source. In addition, it is also motivated by the natural limitations of the point data and grid data: a sparse sensor monitoring network and the low resolution of satellite images. For example, while point sensor data provide high spatial resolution at specific locations, the grid data provide larger spatial coverage but at a lower resolution because they represent an average over a cell of the grid. This trade-off means that while grid data help capture large-scale patterns, they may miss finer-scale variations in soil moisture. By integrating these two data sources, the fused method aims to improve spatial resolution, leading to more accurate and robust soil moisture predictions.

There are many previous studies focusing on data fusion methods, and the following will highlight some of the most relevant research that has contributed to the INLA-SPDE data fusion method. The INLA-SPDE approach is particularly well-suited for modelling complex spatial dependencies and providing accurate predictions for unobserved locations. It can handle large datasets efficiently and is often used in environmental and ecological modelling. A comprehensive review of data fusion methods from a broader perspective can be found in the literature review in Section 1.3. Yang et al. (2023) compare the performance of INLA-SPDE and Random Forest (RF) for soil organic matter (SOM) mapping, using three remote sensing (RS)-based soil moisture indices (NSDSIs) and six Fourier Transform Decomposition (FTD) variables. RF builds multiple decision

trees and aggregates their outputs to improve prediction stability and accuracy. Their results show that INLA-SPDE achieves higher prediction accuracy than RF.

There are many challenges in data fusion modelling, for example, many hydrological variables such as soil moisture, precipitation, and air temperature are continuous across space and over time, but can only be measured at a limited number of locations and time points. In geostatistical approaches such as Kriging, observed data (e.g., point-referenced measurements) are treated as continuous realisations of a spatially correlated random field to obtain interpolation to unobserved locations. However, these methods assume that the observation is continuous in the domain and predictions and observations have the same spatial resolution. This assumption will not hold when fusing different types of data, such as high-resolution point measurements and lower-resolution grid data, leading to a support problem (CoSP). For example, Kriging models trained on point soil moisture data may not represent grid-level variations enough due to aggregation biases, while grid models might smooth fine-scale spatial patterns. Balancing these discrepancies is important for robust data fusion methods because different data sources introduce scale-dependent errors in predictions.

Existing studies of spatial-only INLA-SPDE methods to address CoSP are still limited. To address the CoSP, Moraga et al. (2017) propose a method to combine the point data and grid data using the INLA-SPDE method, mapping observations to a Gaussian Markov Random Field (GMRF) via a novel projection matrix. However, it is a spatial-only model which does not consider the temporal dimension. McMillan et al. (2010) develops a spatio-temporal data fusion model using grid data and presumes that the data process at the point level is linked to the same latent process at the grid level, but it does not consider misaligned covariates. Villejo et al. (2023) develops a two-staged data fusion approach but does not treat the satellite data as block data; instead, they use the centroid of the pixels and treat the grid data as point data. He and Wong (2024) proposed a method to address the CoSP within the INLA-SPDE framework, but does not consider the misaligned covariate and considers the covariate as fixed effects instead of continuous latent fields. These limitations highlight a gap in spatial data fusion, especially in modelling covariates as latent fields while considering misalignment.

This chapter bridges gaps in spatial data fusion models by developing a model that treats covariates as latent fields and considers misaligned covariates, to solve the spatial discrepancies between in-situ sensor data and satellite data using the Integrated Nested Laplace Approximation (INLA) framework. Unlike existing models mentioned earlier in the chapter, which don't consider covariate misalignment or simply treat covariates as fixed effects, this method models covariates as continuous latent spatial processes and considers the misalignment of the covariates. This ensures that fine-scale spatial patterns, which are important for real data applications such as soil moisture mapping, are considered to reduce biases caused by misaligned spatial supports. The chapter begins with methodology and data visualisation, followed by two simulation studies to

validate the fused model under different scenarios, continues with a real-world application to Elliot Water soil moisture data, demonstrating its application in improving spatial resolution in soil moisture map, and ends with a conclusion to discuss the advantages and disadvantages of the model. To assess the models performance, this chapter systematically compares prediction performance (using root mean square prediction error (RMSPE)) for three models: a point model (SEPA In-situ data), a grid model (Copernicus satellite data), and a joint model (SEPA In-situ data and Copernicus satellite data) across different scenarios. This comparison assesses the model's ability to tackle spatial discrepancies.

4.2 Methodology

The spatio-only data fusion model is built upon a geostatistical framework, which is defined as follows. The model assumes that there is a spatially continuous variable underlying all observations that can be modelled using a Gaussian random field process. Let D denote the set of points with real number coordinates in a two-dimensional plane. The process is denoted by $\mu = \{\mu(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^2\}$, has mean function $E[\mu(\mathbf{s})] = 0$ and stationary covariance function $\mathrm{Cov}(\mu(\mathbf{s}), \mu(\mathbf{s}')) = \Sigma(\mathbf{s} - \mathbf{s}')$. Conditionally on $\mu(\mathbf{s})$, point data Y_i observed at a finite set of sites, say $\mathbf{s}_i \in D, i = 1, 2, \ldots, I$, are mutually independent with

$$Y(\mathbf{s}_i) \mid \mu(\mathbf{s}_i) \sim N\left(x(\mathbf{s}_i) + \mu(\mathbf{s}_i), \tau^2\right),$$

where $x(\mathbf{s}_i)$ represents the large-scale structure of the spatial process, capturing broad variation across the study domain. Rather than assuming a constant mean, $x(\mathbf{s}_i)$ allows the expected value of the response variable to vary with location $\mathbf{s}_i$, reflecting the influence of spatially distributed covariates such as elevation. It can be interpreted as a smoothly varying surface that describes the average behaviour of the response variable across space using a mean function, while the remaining spatial random field accounts for variation not explained by the mean function. The τ represents the standard deviation, measuring how much the observations are spread out around the mean.

The geostatistical model framework can be defined as follows for point data and grid data, respectively.
Point data observations in location $\mathbf{s}_i, i = 1, 2, \ldots, I$. Areal data observations are defined as block averages in blocks $\mathbf{B}_j \subset D, j = 1, 2, \ldots, J$, while a block $\mathbf{B}_j$ is a measurable subset of D with $|\mathbf{B}_j| > 0$, over which spatial processes $x(\mathbf{s})$ and $\mu(\mathbf{s})$ are averaged (Moraga et al., 2017).

$$Y_k^{(p)}(\mathbf{s}_i) = \alpha + x(\mathbf{s}_i) + \mu_k(\mathbf{s}_i) + e_k^{(p)}(\mathbf{s}_i), \quad i = 1, \ldots, I \tag{4.1}$$

$$Y_k^{(g)}\left(\mathbf{B}_j\right) = \left|\mathbf{B}_j\right|^{-1} \int_{\mathbf{B}_j} (\alpha_k + x(\mathbf{s}) + \mu_k(\mathbf{s}))d\mathbf{s} + e_k^{(g)}(\mathbf{B}_j), \quad \left|\mathbf{B}_j\right| > 0, \tag{4.2}$$

where $k = 1, 2, \ldots, K$ denote the index for K different variables (such as environmental factors) and B_j denotes a block in domain D and $\left|\mathbf{B}_j\right| = \int_{\mathbf{B}_j} 1 d\mathbf{s}$ denotes the area of $\mathbf{B}_j$, and $e_k^{(p)}(\mathbf{s}) \sim N(0, {\tau_k^{(p)}}^2)$ and $e_k^{(g)}(\mathbf{B}) \sim N(0, {\tau_k^{(g)}}^2)$ are uncorrelated error terms defined by a Gaussian white-noise process $e_k \sim N(0, \tau_k^2)$.

The projection matrix $\mathbf{A}$ specified in the SPDE approach is designed to deal with point-referenced data, and many past studies treat grid data as point data by using the centroid locations of the grid, thereby overlooking the inherent characteristics of grid data. This novel way to construct the projection matrix $\mathbf{A}$ is proposed by Moraga et al. (2017), which specifies that a particular observation in area $\mathbf{B}$ and the process μ is linked through the mean value of the random field in the entire area: $\mu(\mathbf{B}) = |\mathbf{B}|^{-1} \int_{B_j} \mu(\mathbf{s}) d\mathbf{s}$, where $|\mathbf{B}|$ denotes the area of $\mathbf{B}$. The integral defines the theoretical relationship between the latent field $\mu(\mathbf{s})$ and the areal observation $Y_k^{(g)}\left(\mathbf{B}_j\right)$, and represents the true unobserved block average over the entire area B_j. However, computing the integral of $\mu(\mathbf{s})$ is challenging because it is a continuous process, but the mesh vertices used here only have discrete points. The details of the projection matrix and SPDE approach can be found in Section 3.3.3, and the scenario exemplifying the use of barycentric coordinates is shown in Figure 3.2. In the projection matrix, each row of $\mathbf{A}$ corresponds to a particular observation in block $\mathbf{B}_j$. The elements in each row are weights assigned to mesh vertices inside $\mathbf{B}_j$, which are usually 1/H, where H is the number of vertices of the mesh in $\mathbf{B}_j$. So this approximates the integral as $\mu(B_j) \approx \frac{1}{H} \sum_{h=1}^{H} \mu(s_h)$, where s_h are vertices in $\mathbf{B}_j$. Thus, $\mathbf{A}$ acts as a numerical integration operator, converting the continuous integral into a tractable discrete average. The integral of the process in each area is approximated by taking the average of the vertex weights in the corresponding area.

4.3 Simulation study 1: Under the assumption that the point data and grid data have the same measurement errors

Two simulation studies are conducted to assess the spatial data fusion model under different measurement error assumptions. In simulation study 1, it is assumed that both the grid (satellite) and point (sensor) data share the same measurement errors. This allows us to systematically explore questions as follows:

- What is the impact of sensor density on parameter estimation and prediction accuracy, under spatial fields with varying smoothness?
- How well does the joint model handle missing satellite data, such as that caused by cloud cover, and what does this reveal about the models sensitivity to different satellite data

availability?

- How does the spatial resolution of satellite data influence the performance of the joint model, and what does this suggest about the contributions of different sensors and satellites in the fusion process?

Point data observations in location $\mathbf{s}_i, i = 1,2,\ldots,I$ and areal data observations defined as block averages in blocks $B_j \subset D, j = 1,2,\ldots,J$ are defined as follows:

$$\begin{aligned} Y_k^{(p)}(\mathbf{s}_i) &= \alpha_k + \beta_k \times x(\mathbf{s}_i) + \mu_k(\mathbf{s}_i) + e_k^{(p)}(\mathbf{s}_i), \quad i = 1,\ldots,I \\ Y_k^{(g)}(B_j) &= |B_j|^{-1} \int_{B_j} (\alpha_k + \beta_k \times x(\mathbf{s}) + \mu_k(\mathbf{s}))d\mathbf{s} + e_k^{(g)}(\mathbf{B}_j), \quad |B_j| > 0, \quad j = 1,...,J \end{aligned} \tag{4.3}$$

where $k = 1,2,3$ indexes the variables as follows: rainfall (y_1) a covariate that is spatially misaligned with the response variable; soil temperature (y_2) a covariate that is spatially aligned; and volumetric water content (VWC) (y_3) the response variable. The data fusion model developed for soil moisture data uses simulated elevation $x(\mathbf{s}_i)$ as the large-scale structure because elevation is an important predictor of soil moisture and can capture the large-scale trends critical for accurate modelling and analysis. The index i spans both misaligned locations s_i^* $(i = 1,...,n_1)$ and aligned locations s_i $(i = n_1 + 1,...,I)$, which assumes observations exist at both s_i (aligned) and s_i^* (misaligned) locations, with the same linear structure applied to all locations. All the terms in the equation share the same spatial dependency and error structure across all i, which means the aligned covariate and misaligned covariate have the same spatial dependencies and error structure throughout the model. The B_j denotes a block in domain D, $|B_j| = \int_{B_j} 1d\mathbf{s}$ denotes the area of B_j, α_k denotes the intercept, β_k denotes the scaling parameter for the fixed effect, $x(\mathbf{s})$ denotes the large scale trend and $\mu_k(\mathbf{s})$ denotes the latent process.
The prediction value in any unknown locations $\mathbf{s}^p$ within domain D is given by,

$$\hat{Y}_k^{(p)}(\mathbf{s}^p) = \hat{\alpha}_k + \hat{\beta}_k \times \hat{x}(\mathbf{s}^p) + \hat{\mu}_k(\mathbf{s}^p) \tag{4.4}$$

The grid data and the point data measure the same variable, so the latent processes are assumed to be the same. The prediction will be performed on the point data scale using model (4.4).

The main aim of this section is to assess the performance of the data fusion method when combining grid data and point data from a spatial perspective. This section considers several challenging scenarios that are motivated by real datasets, showing the robustness of the fused model. In summary, this section provides a thorough investigation into the effectiveness of the data fusion method under several challenging conditions. The simulation study of this chapter is structured as follows:

- Section 4.3.4 investigates the model's performance across different numbers of locations with latent fields of varying smoothness. This comparison helps to understand the impact

of the sensor density on parameter estimation and prediction performance based on the very sparse sensor monitoring network in the real dataset.

- Section 4.3.5 evaluates the models performance on datasets with varying percentages of missing grid data, simulating the real-world problem of cloud-covering satellite images. These scenarios assess how well the fusion method handles missing data and data gaps, assessing its ability to tackle real-world challenges.

- Section 4.3.6 evaluates how the model performs on a dataset involving grid data of different resolutions, assessing how the fused model handles datasets with different levels of spatial resolution. This gives us insights into the requirements of resolution to make the fused model outperform the point model and grid model, allowing us to understand the relative benefits of fusion approaches when resolutions improve.

4.3.1 Data generation for simulation study 1

The generation of point data in this section follows the methodologies outlined in Section 3.4.3.3. Figure 4.1 shows the simulation process for the data generation at each step. Specifically, the spatial process $\mu(s)$ is modelled through the production of independent realisations from a Matérn Gaussian random field. The first four procedures, which include the generation of the latent field and the values for each variable, are the same as those employed in the simulation study of point data (Section 3.4.3.3). The whole process for simulating data is as follows:

1. The spatial process $\mu(s)$ is simulated by generating independent random field realisations from a Matérn Gaussian random field. The behaviour of the Matérn Gaussian field is controlled through three parameters within the Matérn correlation function: range (ρ), marginal variance (σ), and smoothness (ν).

2. The trend covariate $x(s)$, which represents the geological characteristics of the area, is derived from a surface where values exhibit an increasing pattern from the southwest to the northeast (from 0 to 3.5) across the study area. Let the coordinates of $y(s_i)$ be denoted by Easting_i and Northing_i, then the trend is formulated as follows:

$$x(s_i) = 0.2 * \text{Easting}_i + 0.3 * \text{Northing}_i \tag{4.5}$$

   Additionally, the geographic trend parameter β_3 in Equation (4.3) is defined as -0.2. This may correspond to a surface indicating variations in variables such as soil moisture or other environmental covariates that are associated with changes in latitude and longitude. Figure 3.3 shows the surface of the trend covariate $x(s)$.

3. The uncorrected error terms are generated from a Gaussian white-noise process: $N(0, \sigma_{ek})$.

4. Then the covariates and the response variables are generated by combining the previously constructed terms based on Equation (3.6):

$$
\begin{aligned}
y_1(\mathbf{s}^*) &= \alpha_1 + \mu_1(\mathbf{s}^*) + e_1(\mathbf{s}^*), \\
y_2(\mathbf{s}) &= \alpha_2 + \mu_2(\mathbf{s}) + e_2(\mathbf{s}), \\
y_3(\mathbf{s}) &= \alpha_3 + \beta_3 x(\mathbf{s}) + \beta_1(\alpha_1 + \mu_1(\mathbf{s})) + \beta_2(\alpha_2 + \mu_2(\mathbf{s})) + \mu_3(\mathbf{s}) + e_3(\mathbf{s})
\end{aligned}
$$

Table 4.1 shows the true parameters for the simulation study. The parameters for the simulation study are chosen based on both previous studies and real data characteristics to ensure that they are both theoretically reliable and practically feasible. Some parameters, such as the intercepts α and precision parameters τ, are borrowed from previous studies to maintain comparison with other models. Others, such as scaling parameters β and spatial parameters ρ and σ are chosen from real data applications to reflect spatial patterns and characteristics in the real soil moisture dataset. This allows the simulation to balance theoretical evidence with real data conditions, which makes the assessment of the models performance meaningful.

Table 4.1: True parameter values for the simulation data

	α_1	α_2	α_3	β_1	β_2	β_3	ρ_1	ρ_2	ρ_3	σ_1	σ_2	σ_3	τ_1^2	τ_2^2	τ_3^2
True values	0.5	0.8	1	-0.3	-0.4	-0.2	4	3	2	1	0.5	0.3	0.09	0.04	0.01

5. The grid data is generated by first simulating independent realisations of a Matérn Gaussian random field to model the latent fields. Then, values for each grid cell are computed by averaging all points within that cell, ensuring that each grid cell represents the localised mean of the corresponding latent field. The process can be defined as: $Y^{(g)}(\mathbf{B}) = \frac{1}{n}\sum_{i=1}^{n} y_i$, where $Y^{(g)}(\mathbf{B})$ represents the average value of the grid cell, which indicates the mean of all values y_i within the grid cell. n denotes the total number of points within the grid cell B, y_i represents the value of the ith points within the grid cell.

6. To assess the models ability to generalise unobserved data, the test set includes 20 randomly selected unobserved point locations for the response variable y_3 in each scenario, with the same location across all scenarios in each simulation. Randomly selecting test locations across different simulations gives a comprehensive evaluation of the model's out-of-sample performance, reducing potential bias and assessing how well the fusion model predicts at unobserved locations.

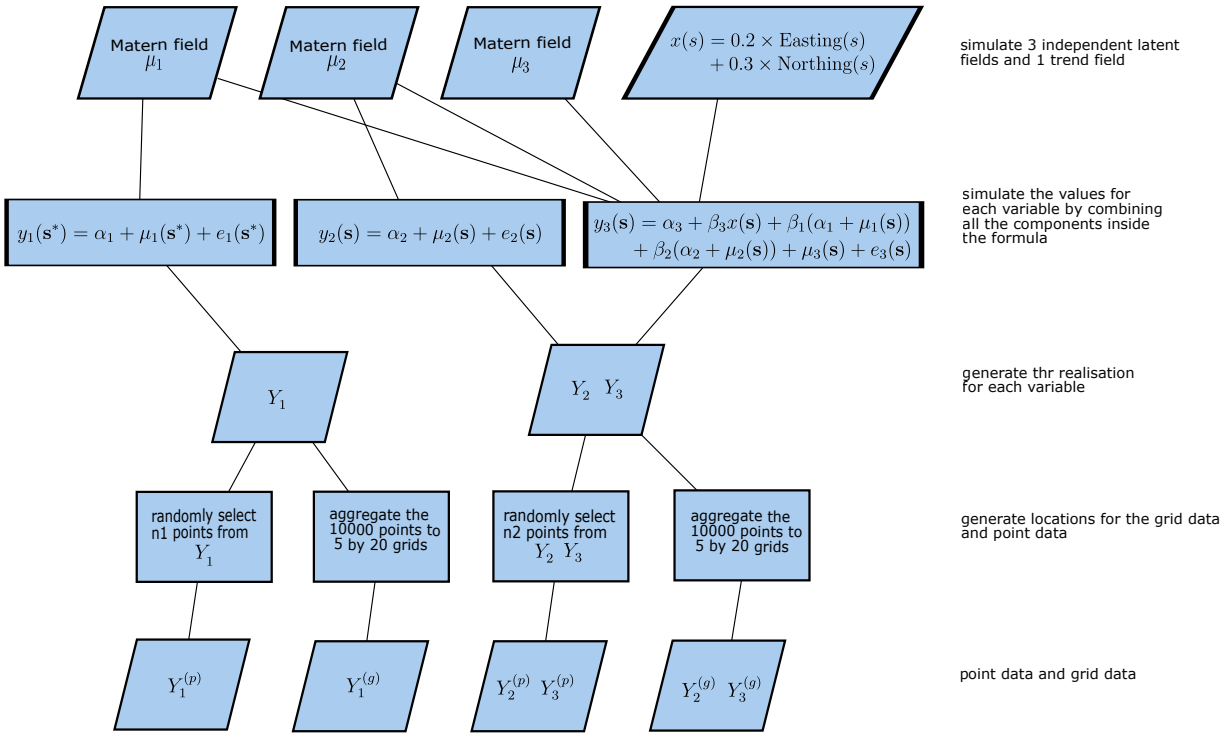
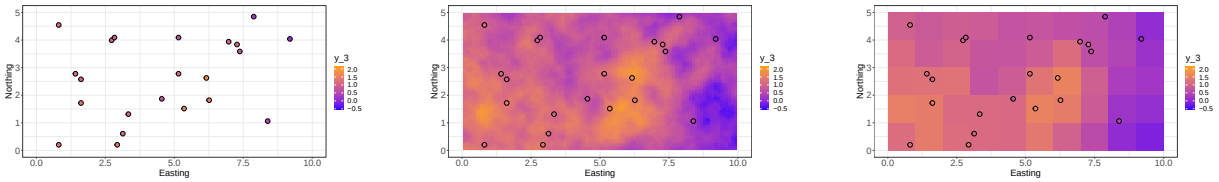


Figure 4.1: Overview of the simulation process, illustrating the generation of three independent Matérn latent fields and one trend field. These components are combined to simulate three spatial variables, which are then used to generate realisations. The realisations are processed into the point and grid data, with point data being randomly sampled and grid data aggregated into a 1×1 resolution.

4.3.2 Simulated latent fields and data visualisation

This section will focus on the visualisations of the latent fields, point data and grid data generated from the above simulation data generation process. The simulation dataset is generated to assess the data fusion model's performance under specific conditions. The latent field models the simulated soil moisture process using a Gaussian Process (GP) with a Matérn covariance function (smoothness parameter $\rho_3 = 2$, variance $\sigma_3 = 0.1$) (Figure 4.1), which captures spatial correlations and allows for random variation. The field is defined over a 5×10 domain with a zero-mean function.

A sensor network is simulated by sampling 22 random locations from the latent field, with additional Gaussian noise ($\tau = 0.1$) accounting for measurement uncertainty (Figure 4.2a). The sensors are randomly distributed over the 5×10 spatial domain. Figure 4.2b presents the true latent field with sensor locations on top of it. The latent field exhibits a clear spatial pattern with variability, showing higher values on the right side of the domain and lower values in the central and left regions. Figure 4.2c) displays the aggregated 1×1 grid data with point data on top of it. Compared to the latent field, the grid data shows a clearer large-scale spatial pattern, as each



(a) Point data (b) Latent field (c) Grid data

Figure 4.2: Simulation data visualisation, with the left panel showing the point data, the middle panel showing the latent field with the point data on top of it, and the right panel showing the grid data with the point data on top of it.

grid cell represents the average of 10,000 simulated samples from the latent field within that grid cell. This aggregation process smooths out local fluctuations, suppressing small random variations while presenting the large-scale pattern of the latent field. This resolution of the grid data is chosen to reflect the characteristics of real satellite images. The spatial parameters, such as the range parameter ρ and grid size, are selected to mimic real data scenarios, which ensures that the simulation reflects the real soil moisture data dispersion and satellite data processing. The simulation parameters of all the simulation data are detailed in Table 4.1.

4.3.3 Simulation priors

The prior setup of the spatio-temporal data fusion model is the same as the prior setup used in Chapter 4. Table 4.2 shows the priors used in the point model, grid model and joint model. The SPDE model will consider the PC-priors (for full details see Section 3.3.4), for the spatial parameters such as the range parameters (ρ_k), and the standard deviation (σ_k^2).

Table 4.2: Priors specification for the joint model parameters.

Parameters	Informative prior	Non-informative prior
α_1		N(0,10)
α_2		N(0,10)
α_3		N(0,10)
β_1		N(0,10)
β_2		N(0,10)
β_3		N(0,10)
ρ_1	$\text{PC}(\rho_0, \alpha)$	
ρ_2	$\text{PC}(\rho_0, \alpha)$	
ρ_3	$\text{PC}(\rho_0, \alpha)$	
σ_1^2	$\text{PC}(\sigma_0, \alpha)$	
σ_2^2	$\text{PC}(\sigma_0, \alpha)$	
σ_3^2	$\text{PC}(\sigma_0, \alpha)$	
$\sigma_{e_1}^2$		$\text{PC}(\tau_{e_1}, \alpha)$
$\sigma_{e_2}^2$		$\text{PC}(\tau_{e_2}, \alpha)$
$\sigma_{e_3}^2$		$\text{PC}(\tau_{e_3}, \alpha)$
$\sigma_{e_1}^2$		$\text{PC}(\tau_{e_1}, \alpha)$
$\sigma_{e_2}^2$		$\text{PC}(\tau_{e_2}, \alpha)$
$\sigma_{e_3}^2$		$\text{PC}(\tau_{e_3}, \alpha)$

The priors for the fixed effects (intercept and slope) and the scaling parameters are normal distributions with a mean of 0 and precision (0.001), which are the default priors with a large variance to ensure the priors provide minimal information. The penalised-complexity (PC) priors are used here for the scale parameter (σ^2) and the spatial variances (ρ) of the Matérn GRFs, with the prior median marginal variance $P(\sigma > \sigma_0) = 0.05$ and the prior median range $P(\rho > \rho_0) = 0.5$ respectively. The PC priors penalise complexity and the distance from the base model by shrinking the range toward infinity and the marginal variance toward zero (Fuglstad et al., 2019). The details of PC priors are in Section 3.3.4. The mean of the standard deviations of y_1, y_2 and y_3, and the mean of ρ_1, ρ_2 and ρ_3 is used as the upper and lower limit of σ^2 and range individually, and the tail probability $\alpha = 0.5$.

4.3.4 Prediction performance between different number of point locations

This section addresses the first research question in Section 4.3.1 by investigating how the model performs under different sensor densities and varying levels of smoothness in the latent spatial field. The analysis aims to evaluate the impact of sensor density on parameter estimation and prediction accuracy, based on a very sparse sensor monitoring network in a real dataset. Predictive accuracy is evaluated using the Root Mean Squared Prediction Error (RMSPE), calculated on a held-out test set. This metric is chosen because it directly reflects the models out-of-sample predictive performance, which is the primary goal in practical applications. Model performance is evaluated by calculating the RMSPE_y across 500 simulations.

The RMSPE_y is defined in Section 3.4.3.8 as follows:

$$\text{RMSPE}_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \quad n = 1, \ldots, 20$$

4.3.4.1 Design of simulation study 1

Table 4.3: Scenarios for evaluating model performance with varying numbers of point locations.

Scenario	y_1	y_2	y_3
Scenario1	10	22	22
Scenario2	20	44	44
Scenario3	40	88	88
Scenario4	80	176	176

Table 4.4: Parameters of the simulated surfaces within scenarios used to assess the impacts of varying marginal standard deviation.

	low variance latent field	medium variance latent field	high variance latent field
α_1	0.5	0.5	0.5
α_2	0.8	0.8	0.8
α_3	1	1	1
β_1	-0.3	-0.3	-0.3
β_2	-0.4	-0.4	-0.4
β_3	-0.2	-0.2	-0.2
ρ_1	4	4	4
ρ_2	3	3	3
ρ_3	2	2	2
σ_1	0.5	1	4
σ_2	0.25	0.5	2
σ_3	0.15	0.3	1.2
$\sigma^2_{e_1}$	0.36	0.36	0.36
$\sigma^2_{e_2}$	0.25	0.25	0.25
$\sigma^2_{e_3}$	0.16	0.16	0.16

In the first experiment, the sensitivity of the data fusion model with misaligned covariates to the number of sensors will be explored. Table 4.3 displays the scenarios used to assess the model's performance with varying numbers of locations. Each scenario specifies the number of observations for: simulated rainfall (y_1) - misaligned covariate; simulated soil temperature (y_2) - aligned covariate; simulated VWC (y_3) - response variable. Scenario 1 represents the true values of the number of different sensors in the real data: $n_1 = 10$, $n_2 = 22$, and $n_3 = 22$, respectively.

Figures 4.3, 4.4, and 4.5 display RMSPE_y values for prediction performance between different numbers of point locations and different smoothness levels, with three violin plots representing

point model, grid model, and joint model. The findings are presented as follows:

Figure 4.4 shows the parameter settings used to generate the simulated surfaces corresponding to low variance, medium variance, and high variance latent fields. For the low-variance latent field, scenario 1 uses the actual number of sensors in the soil moisture dataset. Scenarios 2, 3, and 4 increase the number twofold, fourfold, and eightfold, respectively. At low smoothness, in scenario 1, the median RMSPE_y for the point model is 0.15, higher than the grid model (0.14) and joint model (0.13). When the number of points increases to scenario 4 ($n_1 = 80$, $n_2 = 176$, $n_3 = 176$), the joint model maintains the best performance (median $\text{RMSPE}_y = 0.12$) and the point model (median $\text{RMSPE}_y = 0.11$) outperform the grid model (median $\text{RMSPE}_y = 0.14$).

For the medium variance latent field, in scenario 1, the difference between the grid model and the joint model is small (0.16 vs 0.15), but the median RMSPE_y of the point model (0.26) is noticeably higher than the grid model and joint model. As the number of points increases to Scenario 3 ($n_1 = 40$, $n_2 = 88$, $n_3 = 88$), the point model starts to outperform the grid model, getting a median RMSPE_y of 0.18 compared to the grid model with 0.20. In contrast, the joint model gets the best performance (0.12). The joint model maintains the best performance for all the scenarios, even as the point model gets better performance with the increasing number of points.

For the high variance latent field, the point model ($\text{RMSPE}_y = 0.21$) starts to outperform the grid model (0.23) at scenario 4 ($n_1 = 80$, $n_2 = 176$, $n_3 = 176$), with the best performance belonging to the joint model (0.18). The joint model still maintains the best performance for all the scenarios.

For all levels of smoothness in the latent spatial fields, RMSPE_y values for the point model are higher than for grid and joint models when the number of locations is small, but they will catch up as the number of points increases. To be specific, in a low-variance latent field, the point model needs at least 40 points to outperform the grid model, but in the medium and high-variance latent fields, it needs at least 80 points. When the number of points is small, it is difficult to distinguish differences between the grid model and the joint model. For example, in scenario 1, the median RMSPE_y difference between the grid model and joint model is small (0.16 vs 0.15), but in scenario 4, the difference increases to (0.18 vs 0.10). This suggests that the benefits of the joint model are less significant with fewer data points. But with a larger number of points, for example, in scenarios 2, 3 and 4, the difference between the median RMSPE_y of the grid and joint models becomes more noticeable. This highlights a performance improvement of the joint model as the dataset size increases. Comparing the performance across different levels of latent field variance, the model's performance between the point model, grid model and joint model for all smoothness levels exhibits very similar trends. This comparison helps to understand the impact of the sensor density on the parameter estimation and prediction performance based on the sparse sensor monitoring network in the real dataset, and to determine the minimum sensors

needed to obtain reliable results by varying the number of sensors.

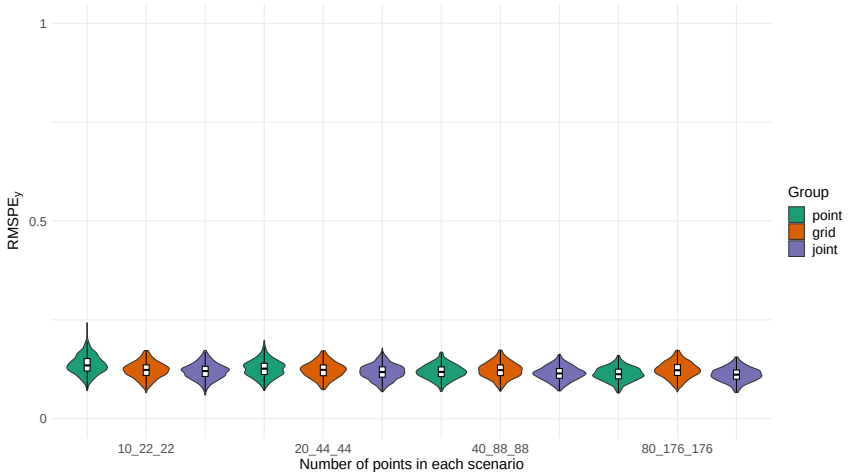


Figure 4.3: RMSPE$_y$ for the point model, grid model and joint model in 500 simulations with low variance latent field ($\sigma_1 = 0.5, \sigma_2 = 0.25, \sigma_3 = 0.15$) for simulation study 1.

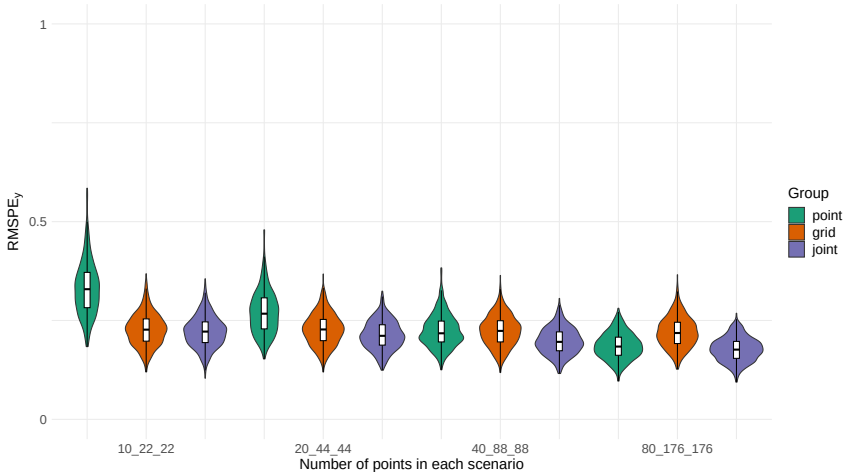


Figure 4.4: RMSPE$_y$ for the point model, grid model and joint model in 500 simulations with medium variance latent field ($\sigma_1 = 1, \sigma_2 = 0.5, \sigma_3 = 0.3$) for simulation study 1.

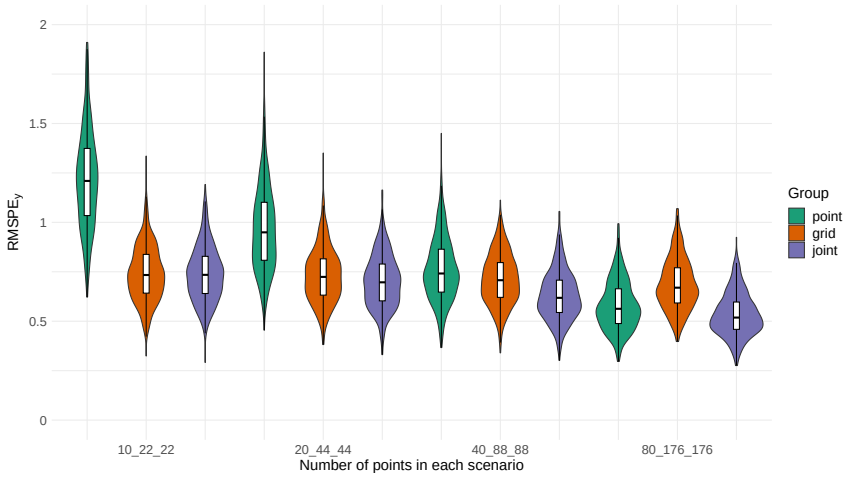


Figure 4.5: RMSPE$_y$ for the point model, grid model and joint model in 500 simulations with high variance latent field ($\sigma_1 = 4, \sigma_2 = 2, \sigma_3 = 1.2$) for simulation study 1.

Table 4.5: The mean of the posterior parameter distributions in the spatial-only model under different numbers of location scenarios.

	True	Scenario 1 (n_1=10, n_2=22, n_3=22)			Scenario 4 (n_1=80, n_2=176, n_3=176)		
		point	grid	joint	point	grid	joint
α_1	0.5	0.5388	0.5250	0.5262	0.5056	0.5248	0.5138
α_2	0.8	0.8000	0.8002	0.7997	0.7995	0.8002	0.7993
α_3	1	1.2879	0.9885	1.1123	1.4142	0.9847	1.5916
β_1	-0.3	-0.1642	-0.0083	-0.0732	-0.2729	-0.0045	-0.4393
β_2	-0.4	-0.2452	-0.0001	-0.0861	-0.3525	-0.0008	-0.4643
β_3	-0.2	-0.2001	-0.1986	-0.1957	-0.1987	-0.1974	-0.2027
ρ_1	4	6.4033	6.2475	5.1610	4.1516	4.3607	4.9935
ρ_2	3	5.6043	3.7910	4.4548	3.7673	5.3291	4.2852
ρ_3	2	5.4453	5.5904	5.2338	3.3208	4.6492	3.0838
σ_1	1	1.0168	0.1598	0.1901	0.9777	0.1396	0.6021
σ_2	0.5	0.5193	0.0953	0.1610	0.5215	0.0925	0.4064
σ_3	0.3	0.5590	0.4680	0.4548	0.3271	0.4798	0.3164
τ_1^2	0.09	0.0940	0.0727	0.1884	0.1024	0.0780	0.2852
τ_2^2	0.04	0.0249	0.0336	2.1816	0.0385	0.0352	0.0834
τ_3^2	0.01	0.0113	0.0077	0.0169	0.0118	0.0095	0.0135

Table 4.5 demonstrates the parameter estimates of the point, grid, and joint models against their true values. The results indicate that parameter estimation improves as the number of point data locations increases. For example, in the joint model, the estimate for α_3 improves from 1.11 (Scenario 1) to 1.03 (Scenario 4), reducing the bias by 72.7% compared to the true value of 1. Similarly, for σ_3 (true value = 0.3), the point models estimate improves from 0.56 (Scenario 1) to 0.33 (Scenario 4), demonstrating an 88.5% reduction in bias with larger datasets. These results confirm that the joint models fusion framework improves estimation accuracy compared to point or grid models.

4.3.5 Prediction model performance with datasets containing partially missing values in the grid data.

In real-world scenarios, such as satellite images, clouds may obscure part of the satellite view, leading to missing data in certain grids at some time points. In contrast, sensor measurements are not affected because they operate independently of weather conditions. To reflect this condition, the simulation focuses on missingness in the grid data, which is assumed to come from satellite data and keeps the point data complete. The simulation aims to assess the model's performance by introducing scenarios where 90%, 50%, and 20% of the grid data are missing at random. These percentages are selected to represent extreme (90%), moderate (50%), and low (20%) missing data conditions, representing both worst-case and small amounts of satellite data missingness.

4.3.5.1 Integration with INLAs predictive framework

The joint model deals with missing data within INLAs computational framework. INLA will automatically compute the predictive distributions for all missing values in the response variable. To be specific, because the response variable's distribution is part of the model, the missing values can be estimated through their predictive distribution. Given a missing response variable $\mathbf{y}_m$, its predictive distribution is defined as follows:

$$\pi(y_m \mid \mathbf{y}_{obs}) = \int \pi(y_m, \boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}})\, d\boldsymbol{\theta} = \int \pi(y_m \mid \mathbf{y}_{obs}, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}})\, d\boldsymbol{\theta}, \tag{4.6}$$

where $\mathbf{y}_{\text{obs}}$ denotes all the response observation and $\boldsymbol{\theta}$ denotes all the hyperparmeters within the model.

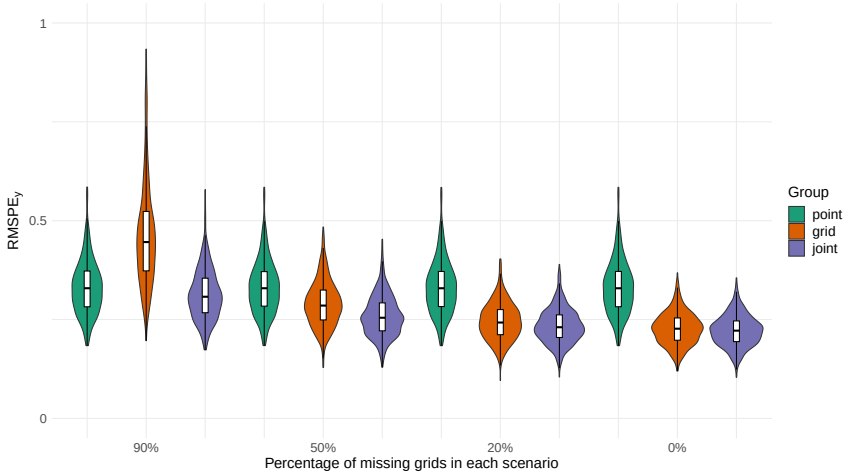


Figure 4.6: Prediction model performance with datasets including partially missing grid data with number of points data (10_22_22) in 500 simulations with medium variance latent fields for simulation study 1.

The model is evaluated under three scenarios with varying percentages of missing grid data: 90%, 50%, and 20%, all of which have the same number of point locations, which is (n_1 = 10, n_2 = 22,

n_3 = 22) and using latent fields with medium smoothness. This simulates real-world conditions where satellite data might be incomplete due to obstructions like clouds. Figure 4.6 displays the RMSPE_y for these different scenarios and suggests that the RMSPE_y grid model decreases as the percentage of missing grid data decreases. For example, the grid model has a median RMSPE_y of 0.40 at 90% missing data, improving to 0.30 at 50%, and 0.20 at 20% missingness. In contrast, the joint model outperforms both the grid and point models across all scenarios, with median RMSPE_y values of 0.25 (90% missing), 0.18 (50% missing), and 0.15 (20% missing). Notably, the joint model's performance stabilises between 20% and 0% missing data, with only a small improvement from 0.15 (20%) to 0.14 (0%), a 6.7% reduction compared to the 37.5% improvement observed between 90% and 20% missingness. The consistency suggests that the joint model is robust to varying levels of missing grid data. Although there is a significant drop between 90% and 20%, it stabilises from 20% to 0%, which suggests that the model is robust when less than 20% of the grid cells are missing. The study demonstrates the robustness of the data fusion model, particularly the joint model, in handling scenarios with significant amounts of missing grid data.

The simulation results from the simulation study 1 have several practical implications for our real data application:

- Minimum sensor numbers: The simulation identifies the minimum sensor numbers needed to get reliable predictions from models as ($n_1 = 40, n_2 = 88, n_3 = 88$) by varying the number of sensors in each scenario. This helps to make decisions on cost-effective sensor deployment without losing prediction and estimation accuracy.

- Latent field smoothness: Different smoothness levels of the latent field affect the model's performance, emphasising the importance of correctly characterising the spatial dependence structure of soil moisture. Understanding this helps design the model to approximate the true underlying environmental processes.

- Robustness to incomplete data: The results show that the joint model maintains stable performance when less than 20% of the grid cells are missing. This suggests that the approach is robust enough to handle low data gaps in real data.

- Effect of grid resolution on prediction performance: This simulation study examines how grid resolution (tested at 1×1, 0.5×0.5, and 0.25×0.25 affects prediction accuracy. At the coarsest resolution, the joint model slightly outperforms the point and grid models, probably because point data compensates for the lack of spatial details. However, at finer resolutions, this advantage disappears, as grid data already captures fine-scale patterns, and the added point data may have limited benefit or slightly reduce performance due to the measurement errors from the point data. These results highlight that the outperformance of joint modelling depends on the resolution of the grid data.

4.3.6 Prediction performance of dataset including different resolution grid data

This section aims to investigate how varying the resolution of grid data affects the prediction performance of the point model, grid model, and joint model. As satellite and remote sensing technology improve, higher-resolution datasets are becoming increasingly accessible, enabling more detailed environmental monitoring and modelling. However, high-resolution data often come with increased computational demands, which leads to practical questions about the optimal resolution for balancing prediction accuracy and computational efficiency. The simulations in the previous section all used grid data with a 1×1 resolution. In this section, three different grid resolutions are considered: 0.25×0.25, 0.5×0.5, and 1×1, corresponding to 800, 200, and 50 grid cells, respectively. To ensure a fair comparison across scenarios, each one includes the same number of point data observations. This helps prevent differences in point data density from impacting the results, allowing for a more focused evaluation of the effects of varying resolution. Figure 4.7 visualises the data for each scenario with the point data on top of the grid data. This section guides how to make decisions on when to choose high-resolution data versus simple, efficient grids.

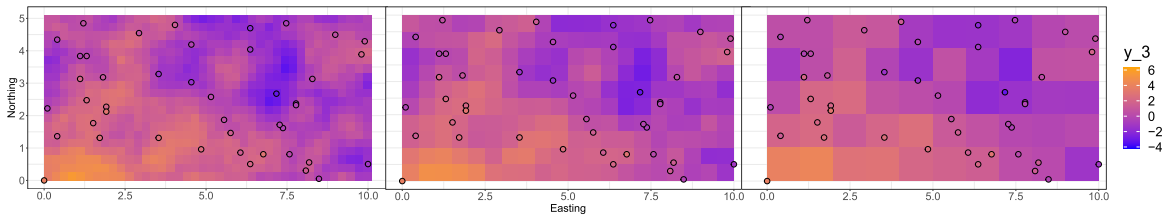


Figure 4.7: Visualisation of different resolution grid data. 0.25×0.25 (Left) 0.5×0.5 (Middle) 1×1 (Right)

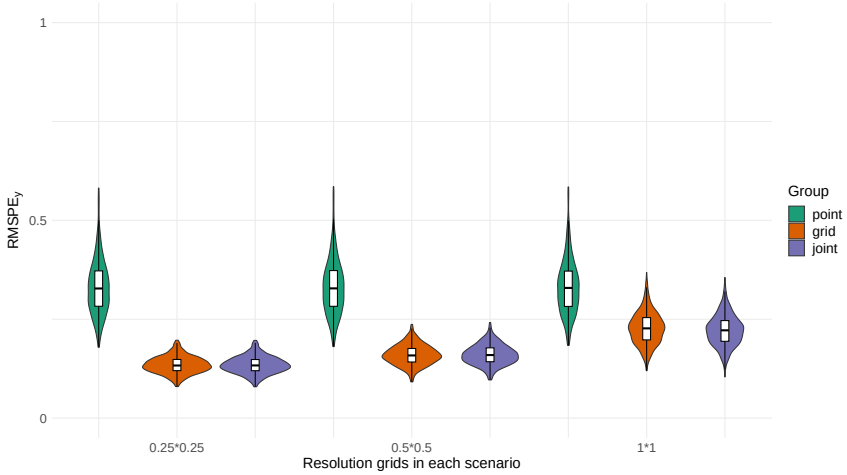


Figure 4.8: RMSPE$_y$ for point, grid and joint models with different grid resolutions: 0.25×0.25 (Left), 0.5×0.5 (Middle), and 1×1 (Right), using point data ($n_1 = 10$, $n_2 = 22$, $n_3 = 22$) across 500 simulations with a medium-variance latent field for simulation study 1.

Figure 4.8 shows the $RMSPE_y$s across datasets with different grid resolutions: 0.25×0.25, 0.5×0.5 and 1×1. The results indicate that the joint model slightly outperforms the point model and grid model when the grid resolution is 1×1. However, this performance of the joint model is not observed for the finer grid resolutions of 0.25×0.25 and 0.5×0.5. The results suggest that the joint model's benefit depends on the resolution of the grid data. To be specific, at the 1×1 resolution, the grid data lacks fine-scale details, and the point data helps compensate by providing additional localised spatial information to fill this gap. At a finer resolution, grid data already capture all fine-scale spatial patterns, so the point data seems redundant and does not contribute to the joint model. In other words, at finer resolutions, grid data may dominate the model fitting and ignore the information contributed from the point data. Point data (with the sensor noise) might harm the model performance when the grid data is already precise. However, the joint model only slightly outperforms the point and grid models in the 1×1 grid data scenario, which could be caused by random variation.

4.3.7 Conclusion

The simulation study systematically evaluates the performance of the data fusion model under different conditions to understand its robustness and performance. The main findings from the simulation study 1 are as follows:

- Effect of sensor density and latent field smoothness (Section 4.3.4):

  Model performance improves consistently with increasing sensor density across all levels of latent field smoothness. At low sensor density, the grid model tends to outperform the point model, particularly for the rough latent fields. However, as the number of sensors increases, the point model catches up and eventually outperforms the grid model, which requires fewer sensors for smoother fields. The joint model consistently achieves the best performance across all scenarios. Improvements in parameter estimation are also observed with increasing sensor density, with bias in parameters such as α_3 and σ_3 reducing by over 72.7% and 88.5%, respectively. This section also helps determine the minimum number of sensors needed for reliable prediction performance under varying spatial smoothness. Specifically, for low-variance latent fields, at least 88 point observations are needed for the point model to outperform the grid model. For medium and high variance fields, the point model requires at least 176 observations. In all cases, the joint model consistently achieves the best performance, with its advantage becoming more pronounced as sensor density increases.

- Impact of missing grid data (Section 4.3.5): The joint model demonstrates strong robustness to missing grid data. While the grid model's performance declines substantially with increasing missingness (e.g., $RMSPE_y$ rises from 0.20 at 20% missing to 0.40 at 90%), the joint model remains stable, with $RMSPE_y$ ranging from 0.15 to 0.25. Notably,

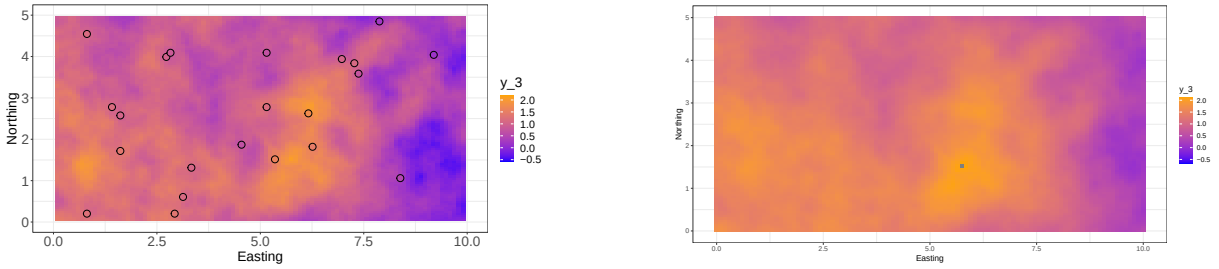
improvements plateau once missingness drops below 20%, indicating that the joint model can effectively compensate for moderate levels of missing grid data. These results highlight the joint model's ability to handle real-world scenarios involving incomplete satellite observations.

- Handling different grid resolutions (Section 4.3.6): The benefit of the joint model depends on the spatial resolution of the grid data. At the coarsest resolution (1×1), the joint model slightly outperforms the point model and grid model, suggesting that point data help fill in missing detail. However, at finer resolutions (0.5×0.5 and 0.25×0.25), the grid data already captures fine-scale variation, making the point data redundant or even leading to slight decreases in performance due to sensor noise. This indicates that the added value of fusing point and grid data diminishes as grid resolution increases.

In summary, the simulation study provides a comprehensive evaluation of the data fusion models advantages and disadvantages, giving some insights into its ability across different spatial resolutions, sensor densities, and data gap scenarios.

4.3.8 Prediction map

Figure 4.9 demonstrates the true latent field and prediction maps of different models with different numbers of point locations in medium variation latent fields (full details are given in Section 4.3.4) to assess the performance of different soil moisture prediction models and compare their predicted spatial patterns against the true latent field. These maps highlight how well each model captures spatial variation and how well predictions align with the ground truth. While all models used the same priors (Table 3.3), parameter estimates (Table 4.5) show posterior means closely aligned with true values. It is noted that the joint model (Figure 4.9d) captures fine-scale variations more accurately than the point model or grid model, suggesting the benefits of the joint model in tackling spatial structure in the environmental process.



(a) True latent field

(b) Point model prediction

(c) Grid model prediction

(d) Joint model prediction

Figure 4.9: True latent field (top left) and prediction maps from the point model (top right), grid model (bottom left), and joint model (bottom right) for the simulation dataset. The true latent field represents the underlying ground truth, while the prediction maps illustrate the estimated values produced by each model across the simulated spatial domain.

4.4 Simulation study 2: Under the assumption that the point data and grid data have different measurement errors

In real-world applications, grid data (e.g., remote sensing data and satellite data) typically have higher measurement errors compared to point data (e.g., ground sensors) due to aggregation biases or resolution limitations. Therefore, to reflect reality, the simulation study will be extended to scenarios where point data and grid data have different measurement errors. This investigation will demonstrate the benefits of the fusion framework in balancing the error conditions from each data stream.

4.4.1 Model specification

While the simulation study 1 in Section 4.3.4.1 assumes the same measurement errors for both point data and grid data, the simulation study 2 relaxes this assumption by introducing different measurement errors for point data and grid data individually. This allows us to investigate how model performance reacts to different noise structures, particularly between point data and grid data. This section builds on the model (4.3) in simulation study 1, with modifications to the measurement structures of the point data and grid data.

Point data observations in location $\mathbf{s}_i, i = 1,2,\ldots,I$ and areal data observations arise as block averages in blocks $B_j \subset D, j = 1,2,\ldots,J$ are defined as follows:

$$\begin{aligned} Y_k^{(p)}(\mathbf{s}_i) &= \alpha_k + \beta_k \times x(\mathbf{s}_i) + \mu_k(\mathbf{s}_i) + e_k^p(\mathbf{s}_i), \quad i = 1,\ldots,I \\ Y_k^{(g)}(B_j) &= |B_j|^{-1} \int_{B_j} (\alpha_k + \beta_k \times x(\mathbf{s}) + \mu_k(\mathbf{s}))d\mathbf{s} + e_k^g(\mathbf{B}_j), \quad |B_j| > 0, \end{aligned} \tag{4.7}$$

where $k = 1,2,3$ denote the index for different variables and B_j denotes a block in domain D, $|B_j| = \int_{B_i} 1 d\mathbf{s}$ denotes the area of B_j, α_k denotes the intercept, β_k denotes the scaling parameter for the fixed effect, $x(\mathbf{s})$ denotes the large scale trend and $\mu_k(\mathbf{s})$ denotes the latent process. The measurement errors $e_k^p(\mathbf{s}) \sim N(0,\tau_1^2)$ and $e_k^g(\mathbf{B}) \sim N(0,\tau_2^2)$ with prior constraints $0 < \tau_1 < \tau_2$ to reflect the higher grid data uncertainty. The grid measurement error $e_k^g(\mathbf{B})$ is used to represent the block measurement error, which means τ_2 represents the total error of the grid data independent of the $|B_j|$. The grid data and the point data measure the same variable, so the latent processes are assumed to be the same. The prediction will be performed on the point data scale using model (4.8). The model (4.7) is the same as model (4.3), except that the point data and grid data have different measurement errors.

The prediction value in any unknown locations $\mathbf{s}^p$ within domain D is given by,

$$\hat{Y}^{(p)}(\mathbf{s}^p) = \hat{\alpha} + \hat{\beta} \times \hat{x}(\mathbf{s}^p) + \hat{\mu}(\mathbf{s}^p) \tag{4.8}$$

4.4.2 Design of simulation study 2

The data are simulated using the model (4.7) with $\tau_1 = 0.1$ and $\tau_2 = 0.3$, maintaining a 1:9 variance ratio to reflect the difference between the point data and grid data (Cressie and Wikle, 2015). There are four scenarios (as shown in Table 4.1) with varying numbers of point locations and the same grid cells for the grid data (1×1 resolution and 50 grids in total). The latent process ($\mu_k(\mathbf{s})$) and the fixed effects use the same priors from Section 4.3.3.

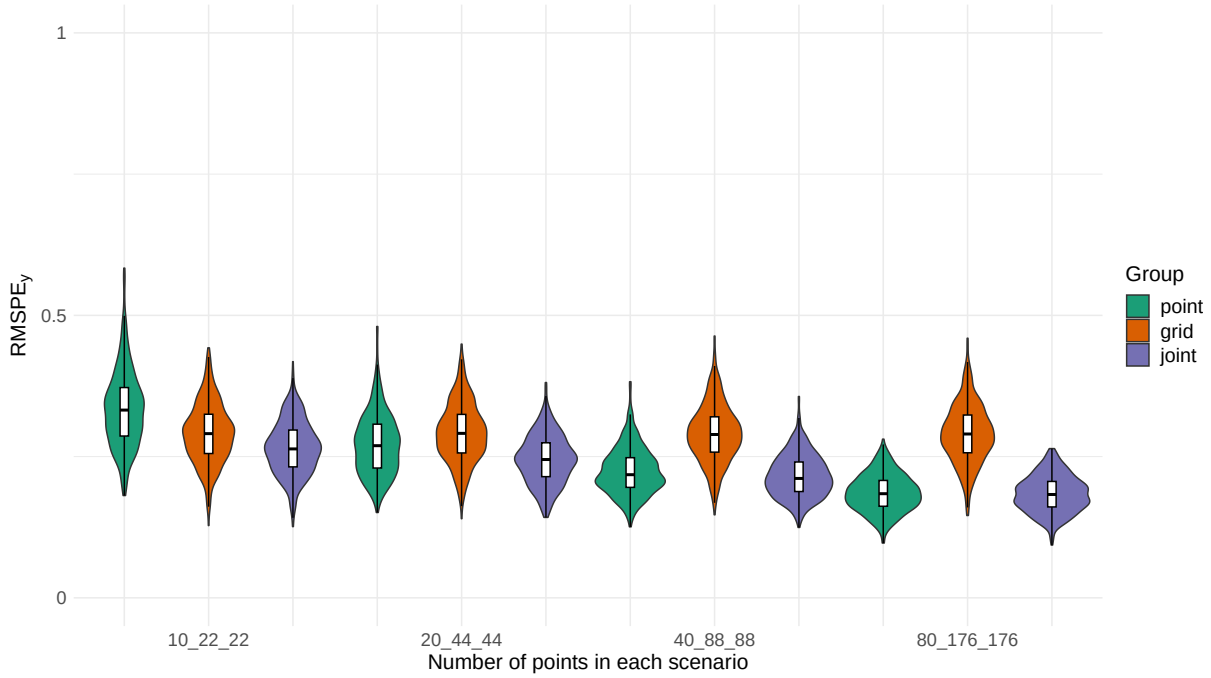


Figure 4.10: RMSPE$_y$ for the point data, grid data and joint data in different scenarios in 500 simulations with medium variance latent field for simulation study 2.

Figure 4.10 compares the RMSPE$_y$ across the point model, grid model and joint model with medium variance latent field. Compared to the results under the same measurement error conditions shown in Figure 4.5, the point model begins to outperform the grid model at an earlier number of point data levels: at (20_44_44) rather than (80_176_176). This highlights that these models are sensitive to the measurement error levels, which suggests that the relative model performance can vary depending on the measurement error levels in the data. The joint model consistently outperforms both the point model and grid model, reducing by around 0.1 across all scenarios. The advantages of the joint model become more pronounced as the number of points increases, since the point data, characterised by lower measurement errors, has a stronger influence on the latent process. The findings show that the benefit of the joint model comes from the high-frequency point data and grids with more spatial coverage (but noisy). When conducting data fusion modelling, it is very important to fully consider the error structure, as this can strongly change the trade-offs between methods. The next section will extend the analysis to a spatio-temporal model, incorporating different levels of measurement error for grid and point data to better reflect real-world conditions. By considering different measurement errors, the model can provide a robust data fusion method in complex multi-source data settings.

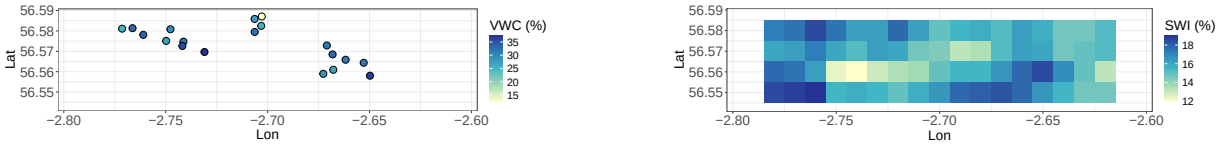
4.5 Real data application

In this section, the real soil moisture data are used to investigate the performance of model (4.7) for real data application. Information on soil moisture in Elliot Water is available as direct

measurements at monitoring site locations (VWC) and as measurements derived from satellite images at a raster grid (SWI). The sensor data have been obtained from 22 sensors located alongside the river, and the satellite image covers the whole study area. The original sensor data are recorded every 15 minutes but are used as daily measurements in this study (a full description is given in Section 2.1). The original measurements of VWC and SWI are displayed in Figure 4.11a and 4.11b. The date 06/05/2021 was chosen for the real data application because both sensor and satellite data were available with good spatial coverage. On this day, 19 sensors were working, which was enough for the analysis. This date also falls within a period when the sensors were performing well and stably. In addition, the rainfall sensors gave consistent readings before and after this date, making the data more reliable for evaluating the fusion model. As the two datasets measure soil moisture in different units, both are standardised using the Z-score transformation to ensure comparability. The covariates *rainfall*, *soil temperature*, and *elevation* may have very different variances, so they are standardised using the Z-score formula as follows:

$$X_{\text{standardised}} = \frac{X - \mu}{\sigma},$$

where X denotes the original value, μ denotes the mean of the variable and σ denotes the standard deviation of the variable. *Rainfall* and *soil temperature* are only available as point-level measurements, and are included as covariates in the point and joint models, but not in the grid model. As a result, only *elevation* is available across all spatial supports and is used as a covariate in all three models.



(a) Raw point data

(b) Raw grid data

Figure 4.11: Point data (19 sensor sites measuring VWC in the Elliot Water) and 95 1km × 1km satellite grid measuring SWI for the same area

The real data application of the data fusion model uses elevation as the large-scale structure because of the availability of elevation, which can capture the large-scale trend in this area. The model is defined as follows:

$$\begin{aligned} VWC^{(p)}(\mathbf{s}_i) &= \alpha_1 + \beta_1 \times rainfall(\mathbf{s}_i) + \beta_2 \times temperature(\mathbf{s}_i) + \beta_3 \times elevation(\mathbf{s}_i) + \mu_3(\mathbf{s}_i) + e(\mathbf{s}_i), \\ SWI^{(g)}(B_j) &= |B_j|^{-1} \int_{B_j} (\alpha_3 + \beta_3 \times elevation(\mathbf{s}) + \mu_3(\mathbf{s}))d\mathbf{s} + e(\mathbf{B}_j), \quad |B_j| > 0, \end{aligned} \tag{4.9}$$

where $i = 1, \ldots, I$, B_j denotes a block in domain D, $|B_j| = \int_{B_i} 1 d\mathbf{s}$ denotes the area of B_j, α denotes the intercept, β denotes the scaling parameter for the fixed effect, and $\mu(\mathbf{s})$ denotes the

latent process. The grid data and the point data measure the same variable, so the latent processes are assumed to be the same. The prediction will be performed on the point data scale using model (4.8).

The prediction value in any unknown locations $\mathbf{s}^p$ within domain D is given by,

$$\hat{Soil\_water\_index}^{(p)}(\mathbf{s}^p) = \hat{\alpha}_3 + \hat{\beta}_3 \times x(\mathbf{s}^p) + \hat{\mu}\left(\mathbf{s}^p\right), \tag{4.10}$$

where $\hat{Soil\_water\_index}^{(p)}(\mathbf{s}^p)$ represents the predicted water index, which is the normalised values of VWC for sensor data and SWI for the satellite data. The $\hat{\alpha}_3$ denotes the estimated intercept; $\hat{\beta}_3$ denotes the estimated effect of the elevation $x(\mathbf{s}^p)$; the latent field $\hat{\mu}\left(\mathbf{s}^p\right)$ captures the dependence structure in the spatial data which is not explained by the covariate. The measurement errors $e^p(\mathbf{s}) \sim N(0, \tau_1^2)$ and $e^g(\mathbf{B}) \sim N(0, \tau_2^2)$ are assumed to follow zero-mean Gaussian distributions, with prior constraints $0 < \tau_1 < \tau_2$. This reflects the assumption through the priors that grid data have higher measurement uncertainty than point data. The priors of the parameters will use the priors described in Section 4.3.3, ensuring methodological alignment.

4.5.1 Leave-One-Out Cross-Validation (LOOCV)

Working with real datasets has a common challenge: a lack of knowledge about the data distribution. Splitting the data into training sets and test sets may not provide enough information about the model's robustness. Cross-validation overcomes these issues by providing a more comprehensive evaluation of the model on multiple subsets of the datasets, providing a more reliable estimation of model performance and how the model will perform on unknown real data.

Given the limited dataset, which includes only 22 sensors, with only 19 available for this particular day (see Figure 4.11), Leave-One-Out Cross-Validation (LOOCV) is used here to fully utilise the real data. LOOCV is particularly suitable for small datasets because it uses each observation for both training data and test data in different iterations. To be specific, for each iteration, one data point is chosen as the test set, while the other 18 points are used as the training set. This process is repeated for each data point to make sure that the model is fully evaluated on every individual point, thus it can provide a fully comprehensive evaluation of the model's performance.

The LOOCV results are visualised from three points of view:

1. Compare actual values with predictions (Figure 4.12).
2. Check the distribution of the residuals (Figure 4.13).
3. Compare the RMSE among the three models (Figure 4.14).
4. Look at how the residuals are distributed across the entire area (Figure 4.15).

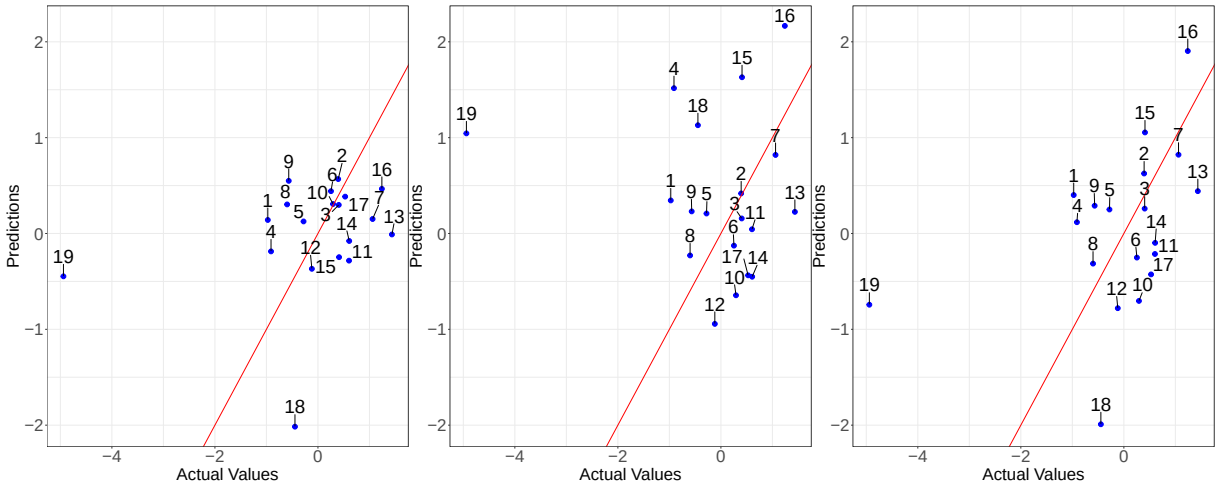


Figure 4.12: Predicted soil water index vs actual standardised soil water index (standardised VWC/SWI) after fitting the following models related to (4.9), with the left panel showing the point model, the middle panel showing the grid model, and the right panel showing the joint model. The red line is the identity line ($y = x$), which represents perfect agreement between the predicted values and actual values.

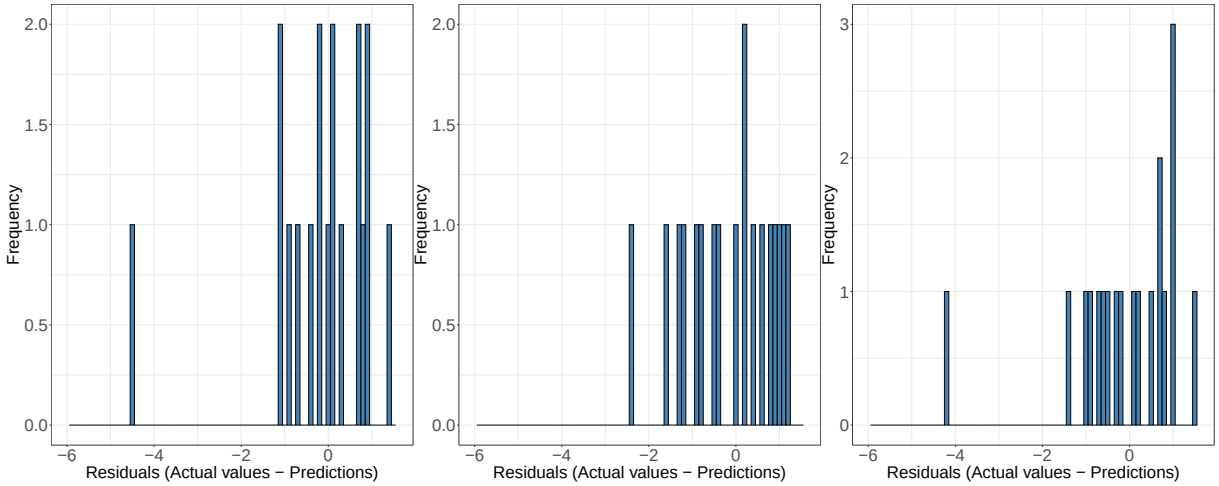


Figure 4.13: The distribution of the residuals after fitting the following models related to (4.9), with the left panel showing the point model, the middle panel showing the grid model, and the right panel showing the joint model.

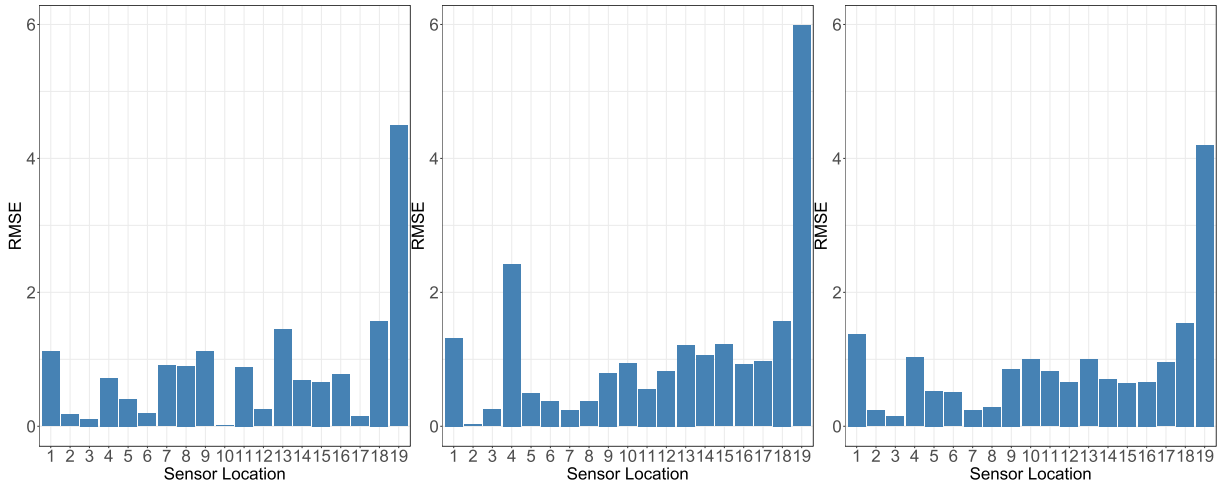


Figure 4.14: RMSE for each test point after fitting the following models related to (4.9), with the left panel showing the point model, the middle panel showing the grid model, and the right panel showing the joint model.

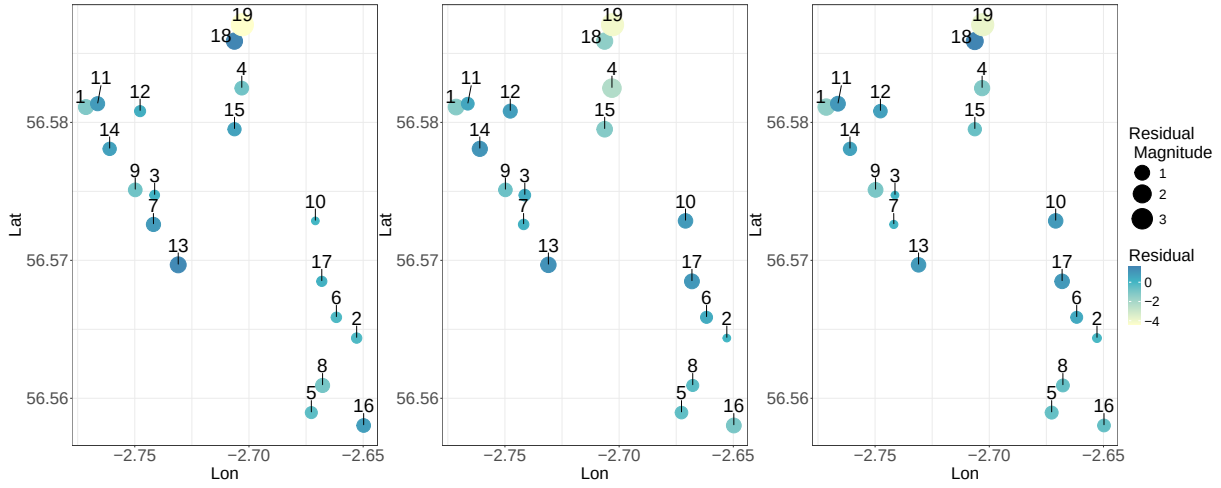


Figure 4.15: Map of the residuals at each location after fitting the following models related to (4.9), with the left panel showing the point model, the middle panel showing the grid model, and the right panel showing the joint model.

Figure 4.12 shows three scatter plots comparing predictions (y-axis) with actual values (x-axis) from the LOOCV results. In the left plot (point model), two sensors (18 and 19) are located next to each other yet exhibit markedly different measurements (as shown in Figure 4.15). This is noteworthy as it violates the spatial autocorrelation law that nearby observations are expected to exhibit similar behaviour. Figure 4.14 shows the RMSE for the point model, where point 19 stands out with much higher RMSE values of 4.49. This is reasonable because sensors typically capture similar values from nearby sites, but for some reason, these two sensors behave quite differently (as shown in Figure 4.11a). Apart from these two outliers, the other 17 points are evenly spread around the line of equality ($y = x$), which suggests that the model performs reasonably well.

In the middle plot (grid model), point 19 is still an outlier, but surprisingly, point 18 is not. This could be caused by the grid, which includes both points, having a value more similar to point 18 rather than to point 19. The other points are spread around the bisector, but the point cluster is not as tight as in the left plot.

Finally, the right plot (joint model) has two outliers: points 18 and 19, and the other points are scattered around the bisector. Figure 4.13 shows the residual distributions for each model. In all cases, the residuals are centred around zero and approximately Gaussian, suggesting that the models have captured the underlying structure well and that no patterns remain in the residuals.

Following the simulation study, the data fusion method is applied to the real-world soil moisture data from SEPA in-situ sensors and Copernicus satellite images. This application aims to validate the models performance under real data conditions, with more uncertainties, sensor noise, and missing data. The analysis will assess the models ability to improve spatial resolution, demonstrating its ability to integrate different data sources for more detailed soil moisture estimation.

4.5.2 Real data prediction results

Point data and grid data often have different spatial scales. For example, point data from sensors is often scattered for design purposes, such as placement alongside a river, whereas grid data usually comes from climate models or radar sensors that generate regular shapes and uniform grids. In the INLA-SPDE framework, the mesh is a discrete representation of the spatial domain used to approximate the continuous Gaussian Random Field (GRF). It plays an important role in the Stochastic Partial Differential Equation (SPDE) approach, which links GRF with Matérn covariance functions to a computationally efficient finite-element representation (the details can be found in Section 3.3.3). Using different meshes for each dataset allows the model to capture most information from each dataset. However, it makes it harder for the model comparison because of the impact of the mesh on the caption of the scale of spatial variation. This section compares the parameter estimation results from models with the same mesh and different meshes to ensure that

the results are both convincing and comparable while capturing the spatial variations properly for each dataset.

The prediction performance of the data fusion model (4.7) is evaluated under two mesh configurations: one where the same mesh is used across all three models (point, grid, and joint), and another where each model uses a mesh constructed based on its respective dataset. The same mesh is constructed using both point data and grid data, which aims to make all the models comparable. The different meshes are constructed using point data, grid data, and joint data individually, which gives an optimisation mesh based on each dataset's location distribution. The meshes are displayed in Figure 4.16, 4.17 and 4.18. The mesh is constructed based on the distribution of the locations, so it can get the optimisation mesh for each dataset. The construction of the mesh needs to deal with the trade-offs between the mesh and the computation time, to be specific, a finer mesh may capture more spatial variation within the data and potentially improve model accuracy, but it may cause huge computational costs and longer computational times (in our experiments, a very fine mesh required roughly five times the runtime of the default mesh used in the main analysis), whereas a coarser mesh can decrease computation time, but it may sacrifice some spatial structure details. The detailed guidelines to get an optimal mesh construction can be found in Section 3.3.3.4.

Figure 4.11a shows that the monitoring sites are located alongside the river, which results in some areas with a high concentration of sites and others with no sites. In the prediction map, the spatial variation is very small because of the sparse distribution of the monitoring sites. The grid data have better spatial coverage, as illustrated in Figure 4.11b, but with a coarse resolution of 1 km. This low resolution results in a less smooth prediction map, so square patterns still appear in the map.

Figure 4.19 shows the prediction maps along with the 95% CIs. The most accurate predictions (narrowest CIs) belong to the joint model, while the largest uncertainty belongs to the point model because of the sparse monitoring network.

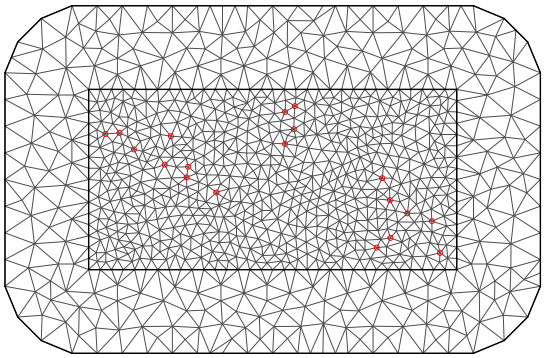


Figure 4.16: Mesh constructed from the spatial distribution of the point data.

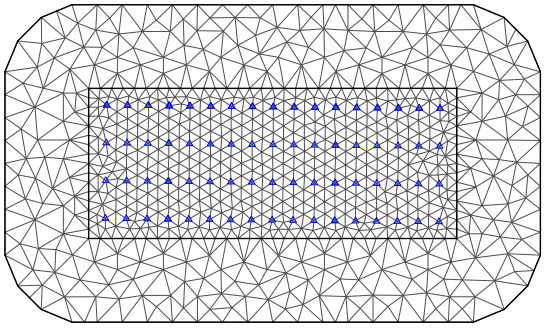


Figure 4.17: Mesh constructed from the spatial distribution of the gridded data.

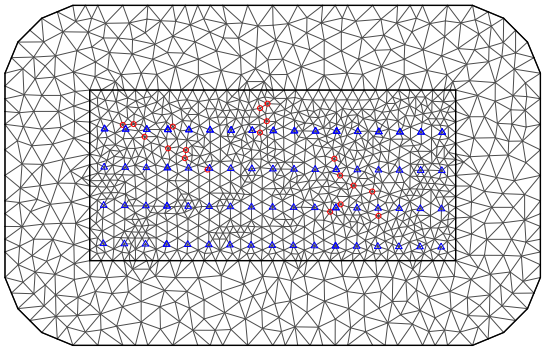


Figure 4.18: Mesh constructed from the spatial distribution of the point and gridded data.

Table 4.6 and 4.7 show the posterior distributions (both point estimates and credible intervals (CIs) for the parameters within the three models, along with PIT (for full details see Section 3.3.5.3) for model checking. In Table 4.6, the same mesh is used for each model. The probability integral transform (PIT) is used to assess the calibration of the model's predictive distribution,

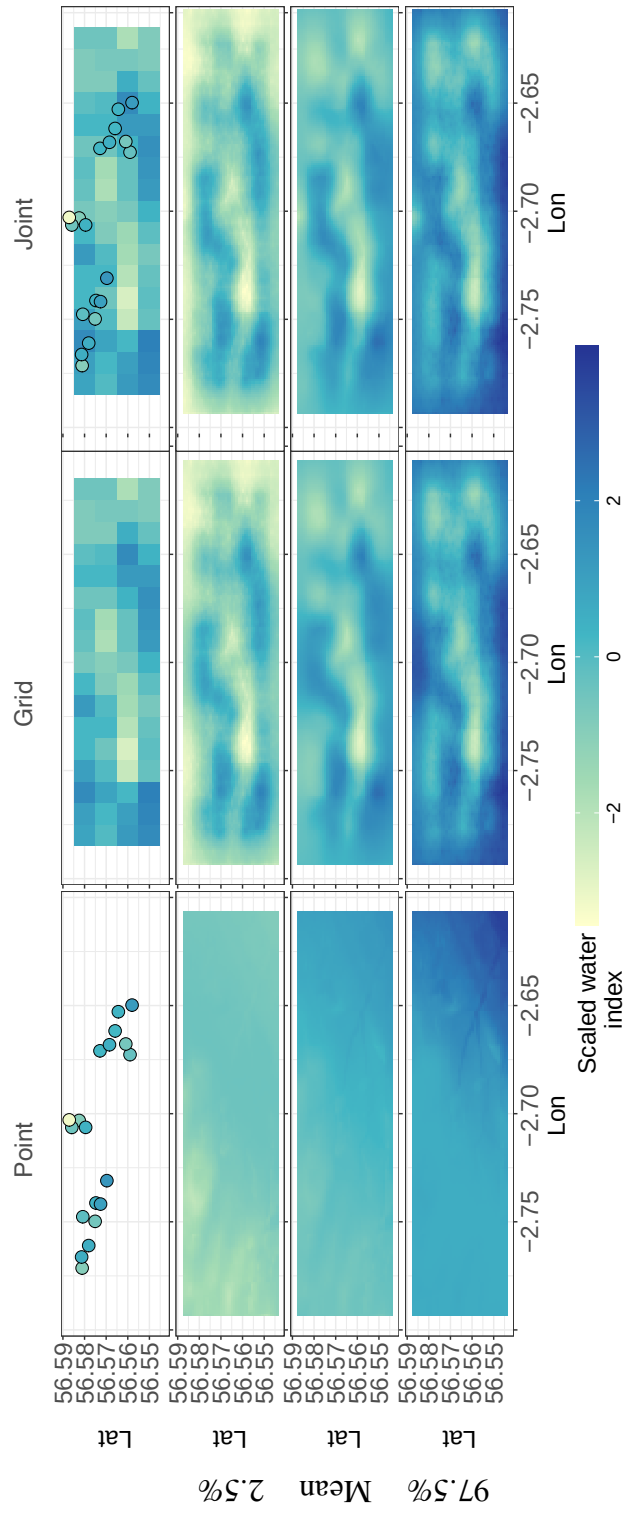


Figure 4.19: Prediction maps of the standardised water index (VWC and SWI) over the Elliott water catchment area, using the same mesh, with 95% credible intervals. The top row displays the real data for each model, while the subsequent rows show the 95% prediction interval, including the 2.5% and 97.5% quantiles as well as the mean of the predictions.

Table 4.6: Parameter estimation with credible intervals for the point model, grid model, and joint model with the same mesh

	Point Model	Grid Model	Joint model
PIT	0.51	0.50	0.50
α_1	-0.09	–	-0.12
	(-0.55, 0.37)		(-0.64, 0.40)
α_2	0.01	–	0.01
	(-0.47, 0.49)		(-0.47, 0.50)
α_3	0.21	0.14	-0.01
	(-0.44, 0.86)	(-0.54, 0.83)	(-0.75, 0.73)
β_1	-0.13	–	-0.12
	(-0.72, 0.46)		(-0.65, 0.40)
β_2	0.02	–	0.11
	(-0.50, 0.55)		(-0.33, 0.55)
β_3	-0.55	-0.19	-0.43
	(-1.45, 0.34)	(-0.77, 0.39)	(-0.99,0.13)
ρ_1	3592.43	–	5492.67
	(251.15, 14400)		(138.77, 30700)
ρ_2	780.67	–	549.39
	(34.10, 3850)		(36.98, 2340)
ρ_3	19179.04	2890	3107.67
	(175.76, 131000)	(1514.89, 5390)	(1587.04, 5850)
σ_1	1.11	–	1.10
	(0.61 1.91)		(0.58, 1.91)
σ_2	0.98	–	1.15
	(0.22, 3.05)		(0.34, 3.31)
σ_3	1.00	1.56	1.66
	(0.32, 2.60)	(0.98, 2.50)	(1.01, 2.72)

Table 4.7: Parameter estimation with credible intervals for the point model, grid model, and joint model with different mesh

	Point Model	Grid Model	Joint model
PIT	0.51	0.50	0.50
α_1	-0.06	–	-0.12
	(-0.71, 0.63)		(-0.65, 0.40)
α_2	0.02	–	0.01
	(-0.45, 0.49)		(-0.47, 0.50)
α_3	0.16	0.22	-0.01
	(-0.42, 0.75)	(-0.77, 1.22)	(-0.75, 0.73)
β_1	-0.06	–	-0.12
	(-0.67, 0.55)		(-0.65, 0.40)
β_2	0.01	–	0.11
	(-0.54, 0.56)		(-0.33, 0.55)
β_3	-0.47	-0.44	-0.43
	(-1.30, 0.36)	(-1.02, 0.13)	(-0.99,0.13)
ρ_1	1407.57	–	5492.67
	(90.21, 7700)		(138.77, 30700)
ρ_2	674	–	549.39
	(35.63, 3060)		(36.98, 2340)
ρ_3	5833.03	3834.97	3107.67
	(100.40, 36200)	(1845.42, 7576.63)	(1587.04, 5850)
σ_1	0.66	–	1.10
	(0.34, 1.16)		(0.58, 1.91)
σ_2	1.43	–	1.15
	(0.68, 2.90)		(0.34, 3.31)
σ_3	1.13	1.80	1.66
	(0.32, 2.95)	(1.01, 3.16)	(1.01, 2.72)

and for a well-calibrated model, the PIT values should follow a uniform distribution on [0,1]. The details of the definition of the PIT can be found in section 3.3.5.3. The mean PIT value is around 0.5, suggesting that the predictive distributions at unobserved locations are well-calibrated and align with the actual observations, indicating reliable prediction performance. However, when examining parameter estimates, particularly the range parameter (ρ_3), the point model yields a different estimate (19179.04; 95% CI: 175.76, 131000) compared to the grid model (2890; 95% CI: 1514.89, 5390) and the joint model (3107.67; 95% CI: 1587.04, 5850). This discrepancy may result from the limited spatial coverage of the point data, which affects parameter estimation even if predictive accuracy remains acceptable. Several sensor sites are close to each other but show different behaviours, which might impact the estimation of the spatial autocorrelation structure. The satellite grid data is 1 km by 1 km, and may not capture the fine-scale spatial variation. The joint model has similar but slightly narrower CIs for most of the parameters, which suggests that the joint model has more robust estimates compared with the grid model. However, the estimates for some parameters β_1 and β_2 have wide intervals, which suggests greater uncertainty in their posterior distributions. This is probably caused by the limited amount of data.

4.6 Conclusion

This chapter evaluates the performance of the spatial-only data fusion model through two simulation studies and an application to real data, explaining how different factors influence parameter estimation and prediction accuracy. These factors include the varying number of point data, the percentage of missing grid cells, the grid data's resolution, and the latent field's variance.

The key findings are as follows:

- Effect of sensor density: For all levels of smoothness in the latent spatial fields, RMSPE_y values for the point model are higher than for grid and joint models when the number of locations is small, but the point model RMSPE_y becomes more similar to the grid and joint as the number of point locations increases. Comparing the performance across different levels of latent field variance, the model's performance between the point model, grid model and joint model for all smoothness levels exhibits very similar trends. However, the advantages of the joint model become more pronounced as the amount of point data increases.

- Robustness to missing grid data: There is a significant drop in performance when the percentage of missing grid observations increases from 20% to 90%, but performance stabilises when the proportion of missing data is below 20%. This suggests that the model is robust as long as more than 80% of the grid data is available.

- Resolution of grid data: The joint model's benefit depends on the resolution of the grid data. At finer resolution, grid data already capture all fine-scale patterns, so the point data

seems redundant and doesn't contribute to the joint model.

- Different measurement errors of point data and grid data: Different measurement errors will help with fusing multi-source data. When conducting data fusion modelling, it is very important to fully consider the error structure, as this can strongly change the trade-offs between methods. In the next chapter, by extending the spatial-only model to the spatial-temporal model, the simulation will consider the different measurement errors.
- Mesh construction: The choice of mesh affects prediction accuracy, computational efficiency, and the integration of point and grid data.

Exploring these scenarios gives a better understanding of the complexities in the spatial-only data fusion. The next chapter extends the spatial-only model to a spatio-temporal model, incorporating temporal dynamics to borrow temporal information.

Chapter 5

Spatio-temporal data fusion model

5.1 Introduction

In chapter 4, a spatial-only data fusion model is developed based on the work of Moraga et al. (2017) for integrating the point data (point-referenced sensor data) and grid data (grid satellite images) while considering misaligned covariates (e.g., the covariates not always observed at the same location of the response variable) and treating the covariates as latent fields. However, a limitation of this spatio-only data fusion method is that the model does not account for the temporal dependence, which is the intrinsic characteristic of many processes. For example, the sensor monitoring network for soil moisture data generates time-point data streams, where an observation at one point can influence future observations (Ochsner et al., 2013b). This dataset serves as the main dataset of the thesis. Ignoring the temporal dependence will reduce the prediction accuracy because the temporal information is not being fully used.

Unlike the spatial-only data fusion model, which relies only on the spatial data from a single day and does not account for the temporal dependencies, the spatio-temporal data fusion model is motivated by the challenge of generating spatio-temporal soil moisture fields when both spatial and temporal data sources are limited in different ways. It aims to address the limitations of the spatial-only method by integrating the temporal information to capture the dynamic nature of the moisture processes. Satellite images provide good spatial coverage but often lack temporal consistency due to cloud coverage. In contrast, in-situ sensors provide high-frequency time series data, but only at a limited number of locations. Rather than simply interpolating in space or time, this chapter aims to learn how soil moisture evolves across both spatial and temporal dimensions by using the rich temporal daily information from both point data and grid data to create a spatio-temporal model that moves from spatial to spatio-temporal dimensions. It is necessary to generate temporally and spatially complete soil moisture datasets, which are essential for water resource modelling, but complete spatio-temporal soil moisture datasets are barely available yet.

This chapter extends the spatial-only model to a spatio-temporal framework with the following

research aims:

- Model extension: extend the spatial-only model to incorporate the temporal dimension, which allows for the modelling of the spatial and temporal variability.
- Model evaluation: evaluate the spatio-temporal model's performance under varying numbers of time points to guide the resource deployments in the study catchment.

This chapter introduces a spatio-temporal data fusion model. It begins with the methodology, followed by a systematic simulation study that evaluates the models performance across varying numbers of time points. The study includes two prediction scenarios: prediction at unknown locations on the last day of the training period, and prediction at unseen locations on a future day. This is motivated by two real-world needs: same-day spatial-gap filling and predicting tomorrow at unknown locations. The chapter continues with a real-world application using soil moisture data from the Elliot Water catchment. This case study combines highly accurate in-situ sensor observations with spatially broad Copernicus satellite imagery to leverage the strengths of both data sources. By integrating point data and grid data while capturing temporal dependencies, the model enhances spatial resolution in soil moisture mapping. The chapter concludes with a discussion of the key findings of the simulation study, along with reflections on its real data application to spatio-temporal data fusion.

5.2 Methodology

The spatio-temporal data fusion model expands the spatial-only data fusion model framework Eq. (4.1) and Eq. (4.2) to include temporal dimensions. While the dimensions are different, the dependence structure between these two models is quite similar. To be specific, the spatio-only process is modelled as a single Gaussian Random Field (GRF) with a Matérn covariance function, a flexible and widely used model for modelling spatial dependence (Stein, 1999). In contrast, the spatio-temporal process is structured as a series of GRFs indexed over time, where temporal dependence is introduced through an AR(1) model on the latent process. This choice is motivated by exploratory data analysis in Chapter 2, which reveals strong short-term autocorrelation in the soil moisture series, and the AR(1) model provides a way to capture such temporal dynamics (Cressie and Wikle, 2015). The spatio-temporal data fusion framework extends the spatio-only data fusion framework to include the temporal dimension. Below, we define the notation for the spatio-temporal latent field that is used to develop the point, grid, and joint models.

5.2.1 Latent field

Let $D \subset \mathbb{R}^2$ denote the spatial dimension and $T \subset \mathbb{R}$ the temporal dimension. The point observations are denoted as $Y^{(p)}(s,t)$. The $\mu(s)$ denotes the spatio-only process, which is a Gaussian

distribution with a Matérn covariance structure that is independent of time. The latent spatio-temporal process is defined as $\eta = \{\eta(s,t) : s \in D, t \in T\}$, with mean function $E[\eta(s,t)] = 0$

We model temporal dependence in the latent field using an autoregressive process of order 1 (AR(1)) based on the exploratory analysis of the real sensor data in Section (2.4.2) :

$$\eta(s,t) = a * \eta(s,t-1) + \sqrt{1-a^2} * \mu(s,t), \quad t = 2,\ldots,M, \tag{5.1}$$

where:

- a as the temporal autoregressive coefficient satisfying $|a| < 1$ to ensure stationarity,
- $\mu(s,t) \sim \text{GP}(0, \Sigma(s,s'))$ is a spatially correlated innovation term defined in Eq. (5.2), which has a Gaussian distribution with a Matérn covariance structure that is independent of time
- initial stage when $t = 0$: $\eta(s,t) \sim N\left(0, \frac{\tau^2}{1-a^2}\right)$.

The covariance function with a Matérn covariance structure is as follows:

$$\text{Cov}_M\left(\mu(\mathbf{s},t'), \left(\mu(\mathbf{s}',t')\right)\right) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\kappa \left\|\mathbf{s}-\mathbf{s}'\right\|\right)^{\nu} K_{\nu}\left(\kappa \left\|\mathbf{s}-\mathbf{s}'\right\|\right), \tag{5.2}$$

where $\|.\|$ denotes the Euclidean distance and K_{ν} is the modified Bessel function of the second kind and ν is the order. To be specific, the modified Bessel function of the second kind is the function $K_n(x)$, which is one of the solutions to the modified Bessel differential equation. The scaling parameter κ can also be interpreted as a range parameter ρ, which represents the Euclidean distance at which $\mathbf{s}$ and $\mathbf{s}'$ become almost independent. The empirically derived definition $\rho = \sqrt{0.8\nu}/\kappa$, corresponds to correlation near 0.1 at the distance ρ, for all ν. The Matérn covariance function $\text{Cov}_M\left(\mu(\mathbf{s},t'), \left(\mu(\mathbf{s}',t')\right)\right)$ defines the dependency structure between two location values measured at the same time point. It is noted that the dependency structure does not account for temporal dependence. Specifically, the covariance between two observations at the same time point is given by

$$\text{Cov}_M(\mu(s,t'), \mu(s',t')) = \Sigma(s-s'),$$

while the covariance between observations at different time points is assumed to be zero:

$$\text{Cov}_M(\mu(s,t), \mu(s',t')) = 0 \quad \text{for } t \neq t'.$$

This implies that the spatio-only latent field $\mu(s,t)$ is correlated only in the spatial dimension and independent over time. The Cov_M denotes the spatial Matérn covariance, but it is noted that we model a day-specific spatial field that is independent across days, while the temporal dependence is handled by a separate term AR(1). Each variable can be assigned its own independent latent process $\eta(s,t)$ following this structure.

Building on the spatio-temporal latent field introduced in Section 5.2, we develop three specific data-fusion models. Model (5.4) links point observations to the latent field. Model (5.6) then extends this framework to the grid data, incorporating gridded remote-sensing data. Model (5.8) fuses point and grid data for joint prediction. Sections 5.2.2 and 5.2.3 introduce the details of the construction of the full spatio-temporal model for point and grid data.

5.2.2 Point level spatio-temporal data fusion model

We model observations at locations s_i, $i = 1,\dots,I$ and time points t_m, $m = 1,\dots,M$, using the following point-level likelihood:

$$Y^{(p)}(s_i,t_m) \mid \eta(s_i,t_m) \sim \mathcal{N}\left(x(s_i,t_m) + \eta(s_i,t_m), \tau_p^2\right), \tag{5.3}$$

where:

- $x(s_i,t_m)$ represents large-scale spatio-temporal covariate (e.g., elevation),
- $\eta(s_i,t_m)$ is the spatio-temporal latent field capturing spatial variation (e.g., rainfall, temperature),
- τ_p^2 is the variance of the measurement errors, quantifying the spread of observations around the mean.

This can be equivalently written as a linear Gaussian regression model:

$$Y^{(p)}(s_i,t_m) = \alpha_k + \beta_k x(s_i,t_m) + \eta_k(s_i,t_m) + e_k^{(p)}(s_i,t_m), \quad i = 1,\dots,I, \quad m = 1,\dots,M, \quad k = 1,\dots,K, \tag{5.4}$$

where:

- α_k is the intercept and β_k is the scaling parameter,
- $\eta_k(s_i,t_m)$ is a zero-mean spatio-temporal latent process for variable k,
- $e_k^{(p)}(s_i,t_m) \sim \mathcal{N}(0,\tau_{pk}^2)$ is Gaussian measurement error.

Full point level spatio-temporal data fusion model:

We define three variables according to our real soil moisture data:

- y_1: rainfall (spatially misaligned covariate),
- y_2: soil temperature (spatially aligned covariate),

- y_3: volumetric water content (VWC, response variable).

Let s_i^* denote misaligned locations ($i = 1,\ldots,n_1$), and s_i denote aligned locations ($i = (n_1 + 1),\ldots,I$). The spatio-temporal fusion model is specified as:

$$\begin{aligned} y_1^{(p)}(s^*,t) &= \alpha_1 + \eta_1(s^*,t) + e_1^{(p)}(s^*,t), \\ y_2^{(p)}(s,t) &= \alpha_2 + \eta_2(s,t) + e_2^{(p)}(s,t), \\ y_3^{(p)}(s,t) &= \alpha_3 + \beta_3 x(s,t) + \beta_1(\alpha_1 + \eta_1(s,t)) + \beta_2(\alpha_2 + \eta_2(s,t)) + \eta_3(s,t) + e_3^{(p)}(s,t). \end{aligned} \tag{5.5}$$

where

- $x(s,t)$ is the spatio-temporal covariate (e.g., elevation),
- $\eta_k(s,t)$ is the spatio-temporal latent field evolves in time via an AR(1) process for variable k,
- $e_k^{(p)}(s,t) \sim \mathcal{N}(0,\tau_{pk}^2)$ is the measurement error specific to variable k,
- β_1 and β_2 quantify the contribution from the two covariates on the response y_3,
- $\eta_3(s,t)$ captures residual spatio-temporal variation in the response variable.

How is the fusion being done?

In this model, spatial misalignment is addressed by incorporating both aligned and misaligned covariates through the latent fields η_1 and η_2. Moreover, the same linear fusion structure is applied uniformly to all spatial locations, no matter whether they are aligned or misaligned. Although each latent process evolves independently over time, they all share a common spatial covariance structure (e.g., Matérn), but with different range and variance parameters. In addition, the covariate $x(s,t)$ (e.g., elevation) captures large-scale spatio-temporal trends in VWC. Finally, the response y_3 fuses latent information from y_1 and y_2, obtaining a flexible and interpretable data fusion framework.

5.2.3 Grid level spatio-temporal data fusion model

For areal observations over blocks $\mathbf{B}_j \subset D$, $j = 1,2,\ldots,J$, while a block $\mathbf{B}_j$ is a measurable subset of D with $|\mathbf{B}_j| > 0$, over which spatio-temporal processes $x(\mathbf{s},t)$ and $\eta(s,t)$ are averaged at time point $t_m : ([t_m,t_{m+1}))$:

$$Y^{(g)}(\mathbf{B}_j,t_m) = \frac{1}{|\mathbf{B}_j|}\int_{\mathbf{B}_j}[\alpha_k + x(\mathbf{B}_j,t_m) + \eta(\mathbf{B}_j,t_m)]d\mathbf{s} + e^{(g)}(\mathbf{B}_j,t_m), \tag{5.6}$$

The notation B_j in the grid model denotes a block in domain D, and $\left|B_j\right| = \int_{B_j} 1\, d\mathbf{s}$ represents the area of block B_j. In the grid model formulation, α_k denotes the intercepts, and $x(\mathbf{B}_j, t_m)$ represents the large-scale structure over the grid block $\mathbf{B}_j$ at time t_m, indicating the average value within that block. The spatio-temporal latent field is denoted by $\eta(\mathbf{B}_j, t_m)$, and the measurement error in the grid data follows $e^{(g)} \sim N(0, \tau_g^2)$. It is important to note that the measurement error of the grid data is assumed to be greater than that of the point data $0 < \tau_p^2 < \tau_g^2$. This relationship is encouraged during model fitting by specifying the prior distribution.

5.2.4 Full spatio-temporal data fusion model

Before introducing the prediction formula, it is noted that the grid-level observations introduced in Section 5.2.3 contribute to the estimation of the second latent field, $\eta_3(s,t)$, and the scaling coefficient $\hat{\beta}_3$.

We define three variables according to our real soil moisture data:

- y_1: rainfall (spatially misaligned covariate),
- y_2: soil temperature (spatially aligned covariate),
- y_3: volumetric water content (VWC, response variable).

Let s_i^* denote misaligned locations ($i = 1, \ldots, n_1$), and s_i denote aligned locations ($i = (n_1 + 1), \ldots, I$). The spatio-temporal fusion model is specified as:

$$
\begin{aligned}
y_1^{(p)}(s^*,t) &= \alpha_1 + \eta_1(s^*,t) + e_1^{(p)}(s^*,t), \\
y_2^{(p)}(s,t) &= \alpha_2 + \eta_2(s,t) + e_2^{(p)}(s,t), \\
y_3^{(p)}(s,t) &= \alpha_3 + \beta_3 x(s,t) + \beta_1 (\alpha_1 + \eta_1(s,t)) + \beta_2 (\alpha_2 + \eta_2(s,t)) + \eta_3(s,t) + e_3^{(p)}(s,t), \\
Y_3^{(g)}(\mathbf{B}_j, t_m) &= \frac{1}{|\mathbf{B}_j|} \int_{\mathbf{B}_j} [\alpha_3 + x(\mathbf{B}_j, t_m) + \eta_3(\mathbf{B}_j, t_m)] d\mathbf{s} + e_3^{(g)}(\mathbf{B}_j, t_m).
\end{aligned}
\tag{5.7}
$$

where

- $x(s,t)$ is the spatio-temporal covariate (e.g., elevation),
- $\eta_k(s,t)$ is the spatio-temporal latent field evolves in time via an AR(1) process for variable k,
- $e_k \sim \mathcal{N}(0, \tau_{pk}^2)$ is the measurement error specific to variable k,
- β_1 and β_2 quantify the contribution from the two covariates on the response y_3,

- $\eta_3(s,t)$ captures residual spatio-temporal variation in the response variable,
- $\left|B_j\right|$ represents the area of block B_j.

In practice, the INLA-SPDE fitting step combines both point and gridded data to produce posterior means of each latent field at the mesh nodes, which are then contributed to prediction locations. The prediction value $\hat{Y_3}^{(pg)}(s,t)$ in any unknown locations s at time t within domain D is given by,

$$\hat{y}_3^{(pg)}(s,t) = \hat{\alpha}_3 + \hat{\beta}_3\, x(s,t) + \hat{\beta}_1\, \hat{\eta}_1(s,t) + \hat{\beta}_2\, \hat{\eta}_2(s,t) + \hat{\eta}_3(s,t). \tag{5.8}$$

Equation (5.8) uses parameter estimates derived from the INLA-SPDE approach to construct predictions for the response variable Y_3. Each item of the prediction model is estimated as follows. The intercept term $\hat{\alpha}_3$ is obtained as the posterior mean of the corresponding fixed effect in the INLA model. The regression coefficient $\hat{\beta}_3$ of the covariate $\hat{x}(\mathbf{s},t)$ is also estimated as a fixed effect using posterior marginals from the INLA-SPDE approach. The scaling parameters $\hat{\beta}_1$ and $\hat{\beta}_2$, which are the weights of the contributions of the latent spatial fields $\hat{\eta}_1(\mathbf{s},t)$ and $\hat{\eta}_2(\mathbf{s},t)$, are similarly estimated as fixed effects within the hierarchical model.

The latent fields $\hat{\eta}_1(\mathbf{s},t)$, $\hat{\eta}_2(\mathbf{s},t)$, and $\hat{\eta}_3(\mathbf{s},t)$ are modelled using the INLA-SPDE approach, which approximates the continuous spatial fields with Gaussian Markov random fields (GMRFs) defined using triangulated meshes. Their posterior means at the mesh nodes are computed during the model fitting step and then projected to the prediction locations through the basis functions of the SPDE mesh. The spatial smoothness, range, and marginal variance of each latent field are treated as hyperparameters, inferred from their posterior distributions, and summarised by posterior means.

The predicted response $\hat{Y_3}^{(p)}(\mathbf{s},t)$ is constructed by combining the estimated intercept, the scaled covariate effect, the weighted contributions from the two latent fields, and the direct latent field representing Y_3. Although predictions can be generated for all variables, the focus here is on evaluating predictive performance specifically for the response variable y_3. In generating predictions, the outputs from the fine-resolution maps are treated as point data.

In summary, we developed three spatio-temporal fusion models above: the point model is fitted using only the point observations $Y^{(p)}(s,t)$, the grid model is fitted using only the gridded (remote-sensing) data contributing to $\eta_2(s,t)$, and the joint prediction model fit to both data sources, integrating them through the latent fields.

5.3 Simulation study

The simulation study aims to evaluate how varying the number of time points improves the accuracy of prediction and parameter estimation of a spatio-temporal data fusion model, motivated by the real-data challenge of integrating two data sources that can compensate for each other but are not perfect by themselves: sparse point soil moisture sensor data (high resolution but poor spatial coverage) and satellite grid data (broad spatial coverage but poor resolution). While the spatial-only model in Chapter 4 has already demonstrated that the joint model outperforms the point model and grid model (full details are given in Section 4.6), extending this framework to a spatio-temporal model can answer important questions. To be specific, the model incorporates temporal dependence, which allows the model to borrow information across both space and time, improving parameter estimation by capturing evolving spatial patterns and temporal trends. Additionally, from a real data application perspective, a spatio-temporal framework can support one-step-ahead prediction, which is important for predicting future soil moisture to support agricultural or environmental management. Finally, the simulation study determines the minimum number of time points required to achieve reliable soil moisture predictions by simulating scenarios with varying numbers of time points. It uses spatio-temporal data from both in-situ sensors and satellite images, offering insights into data fusion strategies for real data applications.

The simulation study section begins by introducing the simulation design, which outlines two scenarios used to evaluate the performance of the spatio-temporal model under different numbers of time points. The simulation design also details the data generation process used in the simulation study. This is followed by a visualisation of the simulated data to illustrate the spatial and temporal patterns across the latent fields, point observations, and grid data. The section concludes with the prediction performance of the spatio-temporal data-fusion model on the simulated data for both scenarios.

5.3.1 Simulation design

To evaluate the predictive performance of the spatio-temporal data fusion model, two simulation scenarios are designed. The first scenario focuses on spatial prediction, where the model is used to predict values at unknown locations on the final day of the training period. This assesses the models spatial interpolation ability, leveraging the temporal information in the training set, and provides a direct comparison with the spatio-only results in Chapter 4. The second scenario focuses on spatio-temporal prediction, where the model predicts values at unknown spatial locations one day ahead, for example, outside the temporal range of the training data. This tests the model's ability to both interpolate in time and generalise to unseen locations.

- **Scenario 1**: Predict on the final day of the training period at previously unobserved locations, using $k = 3, 7, 10$, and 30 time points.

- **Scenario 2**: Do one-day-ahead forecasts beyond the training period at unobserved locations, again for $k = 3, 7, 10$, and 30 time points.

Both of these scenarios provide a comprehensive assessment of the models ability to fuse point sensor data and grid satellite data while maintaining accuracy across different spatial and temporal conditions.

The generation of spatio-temporal data in this section follows the methodology in Section 5.2. Figure 5.1 illustrates the step-by-step simulation process used for data generation, outlining how the latent fields, point data, and grid data are constructed at each stage. Specifically, the spatial process $\mu_k(s)$ is modelled through the production of independent realisations from a Matérn Gaussian random field. The first four procedures are the same as the spatio-only data generation, and then the temporal correlation is introduced by the AR(1) model. The following procedures are the same as the simulation data generation of the spatio-only model (full details are given in Section 4.3.1). The whole process (The flowchart can be found in Figure 5.1) for simulating data is as follows:

1. Spatial processes $\mu(s)$ are simulated by generating 100 independent random field realisations from a Matérn Gaussian random field as fixed seeds, and the same 100 seeds are reused when constructing datasets of length k days. Holding the spatial field fixed over time makes sure that differences across scenarios arise from k, not from spatial variability. The behaviour of the Matérn Gaussian field is controlled through three parameters within the Matérn covariance function: range (ρ), marginal variance (σ), and smoothness (v).

2. The temporal correlation is introduced by the formula as follows:

$$\eta(s,t) = a * \eta(s,t-1) + \sqrt{1-a^2} * \mu(s,t),$$

   where the $\sqrt{1-a^2}$ term is used to make the process stationary in time. The spatio-temporal process is assumed to be a series of GRFs, and the latent spatial processes $\mu_k(s)$ generated from the first step account for the temporal dependencies through AR(1).

3. The trend covariate $x(s,t)$, which represents the geological trend of the study catchment, is derived from a surface where values exhibit an increasing pattern from the southwest to the northeast (from 0 to 3.5) across the whole study catchment. Let the coordinates of $y(s_i,t)$ be denoted by Easting_i and Northing_i, then the trend is defined as follows:

$$x(s_i,t) = 0.2 * \text{Easting}_i + 0.3 * \text{Northing}_i \tag{5.9}$$

   Additionally, the geographic trend parameter β_3 of the trend covariate in Equation (5.4) is defined as -0.2.

4. The uncorrected measurement error terms for point data (e_p) and grid data (e_g) are generated from a Gaussian white-noise process: $N(0,\tau_p^2)$ and $N(0,\tau_g^2)$.

5. Then the covariates and the response variables are generated by combining the previously constructed terms based on Equation (5.4):

$$
\begin{aligned}
y_1(\mathbf{s}^*,t) &= \alpha_1 + \eta_1(\mathbf{s}^*,t) + e_1(\mathbf{s}^*,t),\\
y_2(\mathbf{s},t) &= \alpha_2 + \eta_2(\mathbf{s},t) + e_2(\mathbf{s},t),\\
y_3(\mathbf{s},t) &= \alpha_3 + \beta_3 x(\mathbf{s},t) + \beta_1(\alpha_1 + \eta_1(\mathbf{s},t)) + \beta_2(\alpha_2 + \eta_2(\mathbf{s},t)) + \eta_3(\mathbf{s},t) + e_3(\mathbf{s},t)
\end{aligned}
$$

Table 5.1 shows the true parameters used in the spatio-temporal simulation study. The parameters for the simulation study are chosen based on both previous studies and real data characteristics to make sure that they are both theoretically reliable and practically feasible. Some parameters, such as the intercepts α and precision parameters τ, are borrowed from previous studies to maintain comparison with other models. Others, such as scaling parameters β, spatial parameters ρ and σ, and the temporal coefficients a, are chosen from real data applications to reflect spatial patterns and characteristics in the real soil moisture dataset. This allows the simulation to balance theoretical evidence with real data conditions, which makes the assessment of the models performance meaningful.

Table 5.1: True parameter values used in the spatio-temporal simulation data

	α_1	α_2	α_3	β_1	β_2	β_3	ρ_1	ρ_2	ρ_3	σ_1	σ_2	σ_3	τ_{p1}^2	τ_{p2}^2	τ_{p3}^2	τ_{g1}^2	τ_{g2}^2	τ_{g3}^2	a_1	a_2	a_3
True values	0.5	0.8	1	-0.3	-0.4	-0.2	4	3	2	1	0.5	0.3	0.09	0.04	0.01	0.25	0.16	0.09	0.4	0.5	0.6

6. The grid data is generated by first simulating independent realisations of a Matérn Gaussian random field to model the latent fields. Then, for each time point, values for grid cells are calculated by averaging all points within that cell to ensure that each grid cell represents the localised mean of the specific latent field. The process can be defined as: $Y^{(g)}(\mathbf{B}) = \frac{1}{n}\sum_{i=1}^{n} y_i$, where $Y^{(g)}(\mathbf{B})$ represents the average value of the grid cell, which indicates the mean of all values y_i within the grid cell, n denotes the total number of points within the grid cell $\mathbf{B}$, and y_i represents the value of the ith points within the grid cell in each day.

7. To assess the models ability to generalise unobserved data, the test set includes 20 randomly selected unobserved point locations for the response variable y_3 on the final day of the training period (Scenario 1 in Section 5.3.1) and one day ahead of the training period (Scenario 2 in Section 5.3.1), with test points always at the same locations across point model, grid model and joint mode within each simulation. Randomly selecting test locations across different simulations gives a comprehensive evaluation of the model's

out-of-sample performance, reducing potential bias and assessing how well the fusion model predicts at unobserved locations.

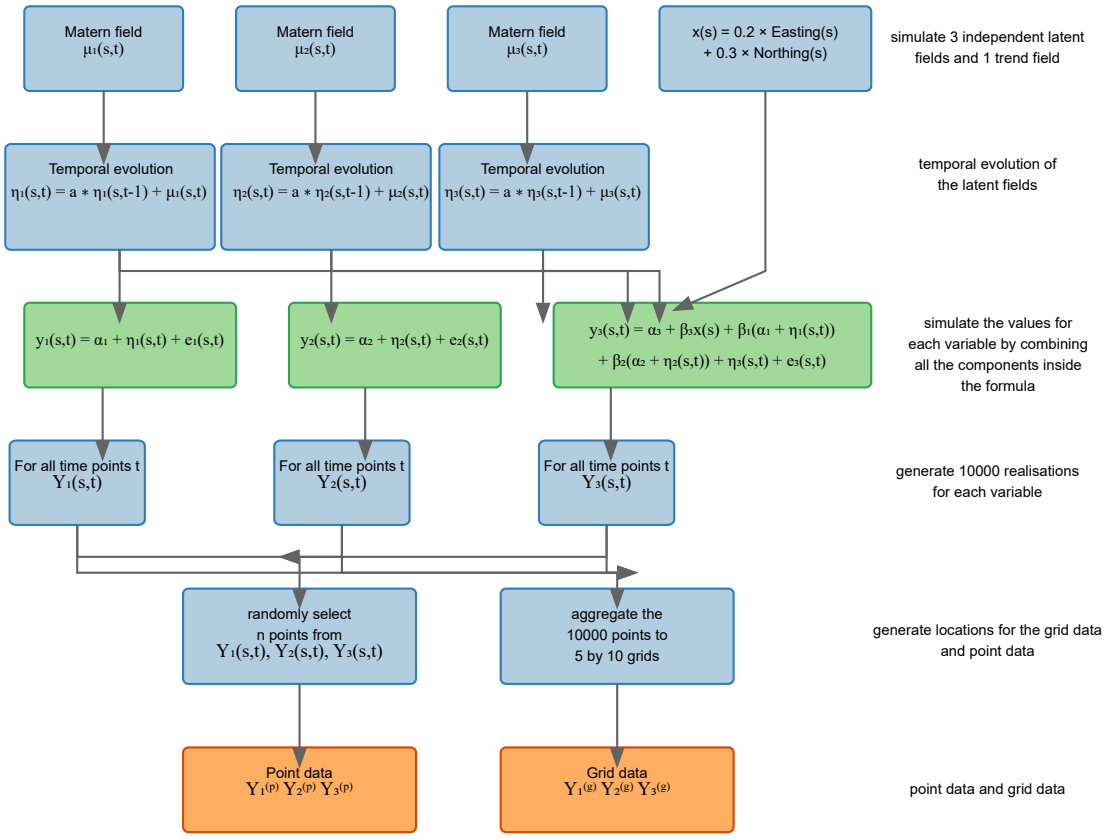


Figure 5.1: Flowchart illustrating the simulation process for the spatio-temporal study. The first row shows the generation of independent latent fields for each variable. The second row introduces temporal dependence. The third row combines these components into a full spatio-temporal realisation. Subsequent rows demonstrate how grid data and point data are derived from this realisation.

5.3.2 Characterisation of point and gridded data

This section characterises the spatio-temporal simulation data used in the spatio-temporal simulation study, which provides insight into how spatial patterns evolve over time.

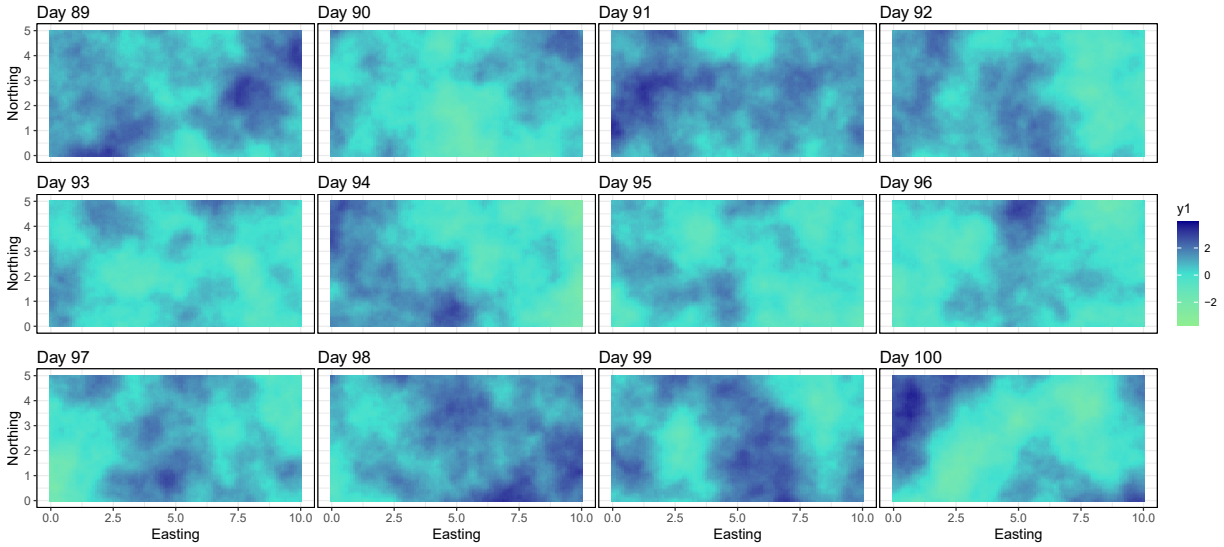


Figure 5.2: Final 12-day sequence of simulated latent fields for simulated rainfall y_1 in the spatio-temporal simulation study, using a medium variance latent field configuration ($\sigma_1 = 1$, $\sigma_2 = 0.5$, $\sigma_3 = 0.3$) and a temporal coefficient $a_1 = 0.4$.

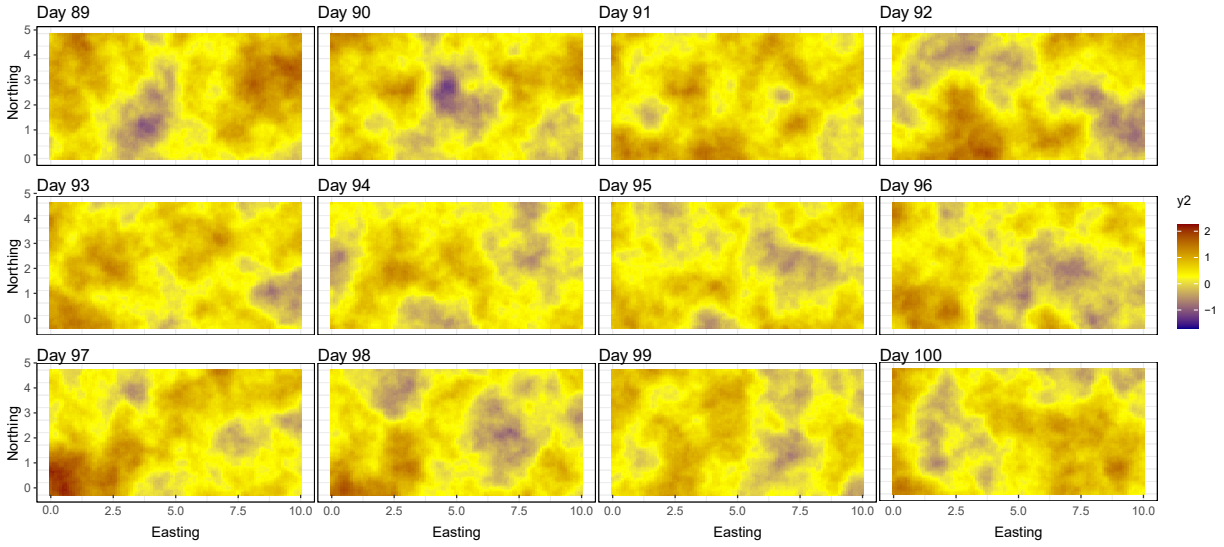


Figure 5.3: Final 12-day sequence of simulated latent fields for simulated soil temperature y_2 in the spatio-temporal simulation study, using a medium variance latent field configuration ($\sigma_1 = 1$, $\sigma_2 = 0.5$, $\sigma_3 = 0.3$) and a temporal coefficient $a_2 = 0.5$.

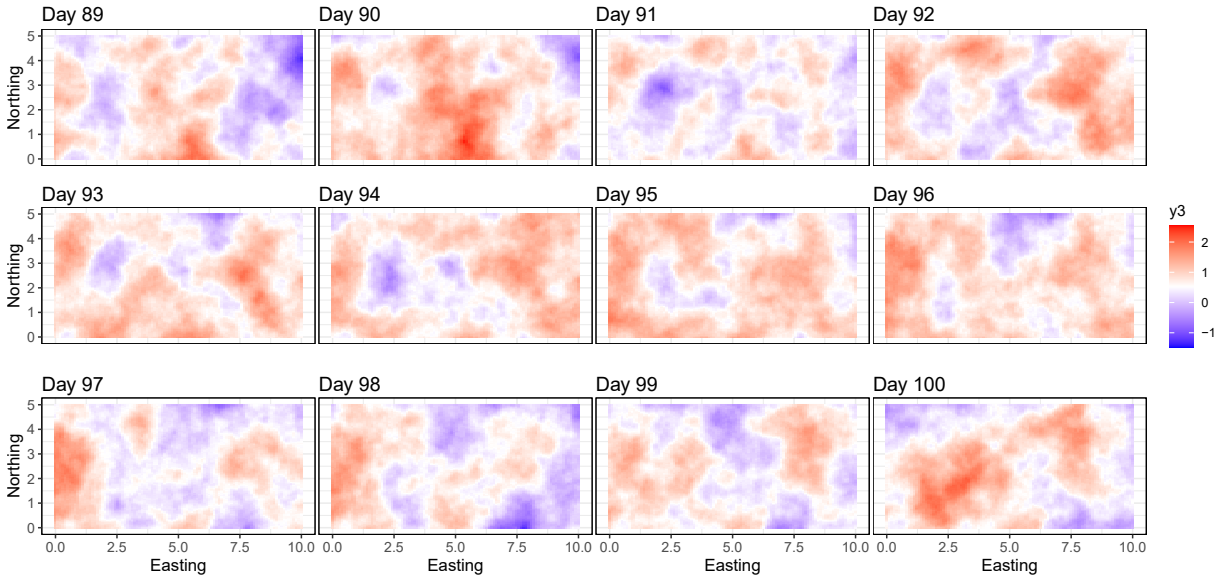


Figure 5.4: Final 12-day sequence of simulated latent fields for simulated soil moisture y_3 in the spatio-temporal simulation study, using a medium variance latent field configuration ($\sigma_1 = 1$, $\sigma_2 = 0.5$, $\sigma_3 = 0.3$) and a temporal coefficient $a_3 = 0.6$.

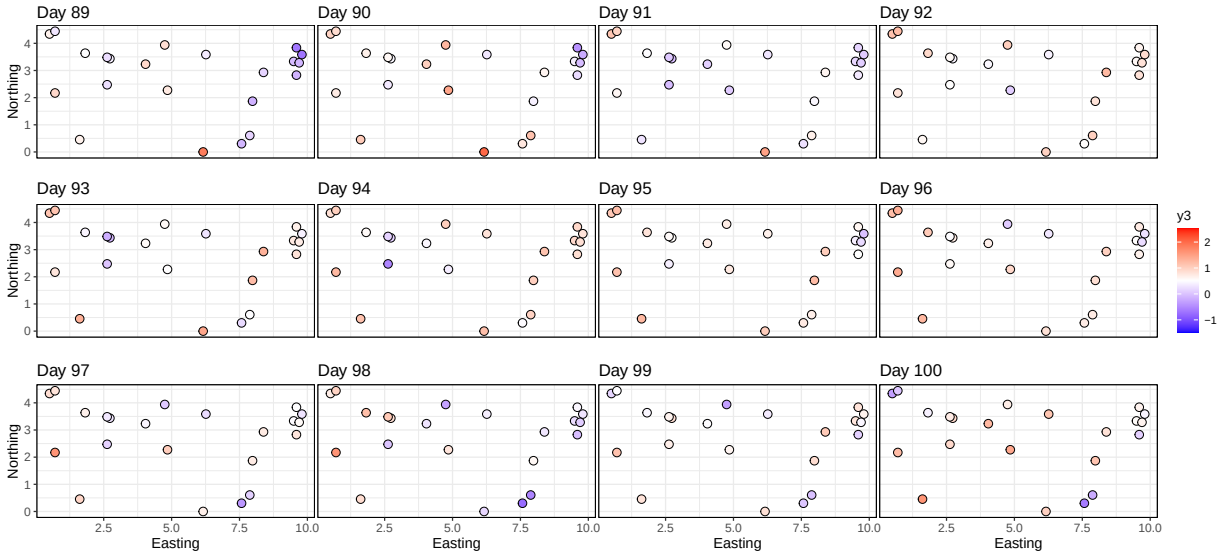


Figure 5.5: Final 12-day of simulated point soil moisture data for y_3 in the spatio-temporal simulation study. Twenty-two points are randomly selected from the realisation surface of simulated soil moisture y_3 on each day.

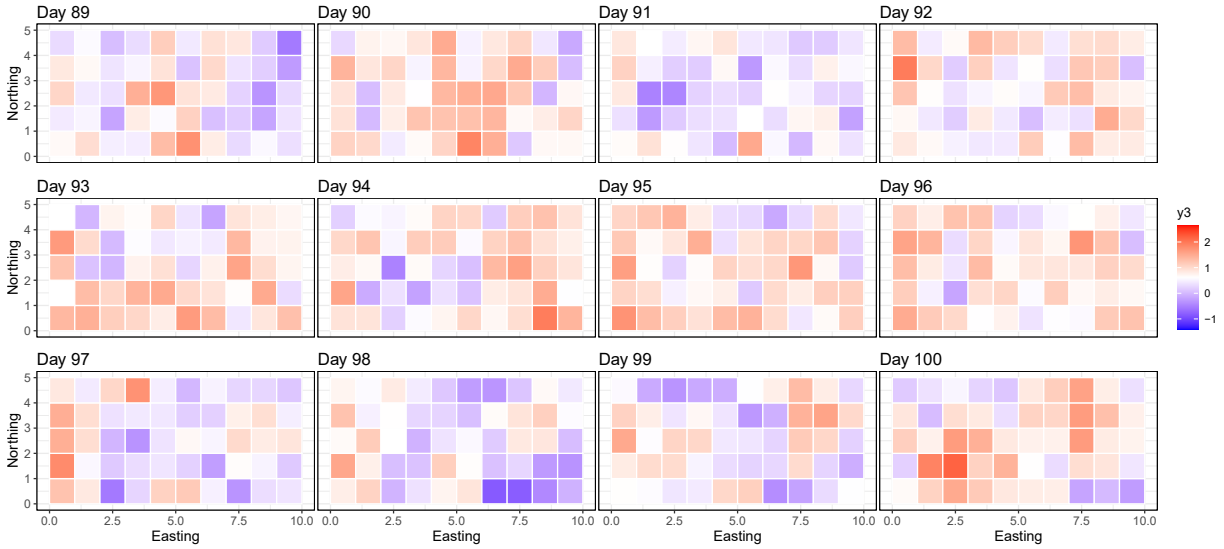


Figure 5.6: Final 12-day of simulated grid soil moisture data for y_3 in the spatio-temporal simulation study. It is averaged by 10,000 points generated from the realisation surface of simulated soil moisture y_3 on each day.

The latent field η_3 simulates soil moisture dynamics using a spatially continuous and temporally autoregressive dependence structure. The spatial field is modelled as a Gaussian process with a Matérn covariance (smoothness $\rho_3 = 2$, variance $\sigma_3 = 0.1$), and the temporal dynamics as an AR(1) process with coefficients $a_1 = 0.4$, $a_2 = 0.5$, and $a_3 = 0.6$ corresponding to rainfall, soil moisture, and VWC, respectively. The autoregressive parameters are chosen based on the exploratory analysis of the real sensor data in Section (2.4.2), which reflects the characteristics of the real data. The simulation fields span over a 5×10 spatial domain and evolve over 100 simulated time points, but only the final 12 are shown here.

Both the point data and grid data, along with the latent fields, are simulated by the model (5.4) and (5.6) according to the workflow in Figure 5.1. The sensor network is simulated by sampling 22 locations across all time points. At each time point, we sample 10,000 values to evaluate the Matérn Gaussian random field, and average the values within each grid cell to approximate the spatial integral. After the independent daily data are generated, the AR(1) process is used to introduce the temporal dependence across time into the data.

The point data include a measurement error denoted as τ_{p_3}. In contrast, the grid data includes a measurement error denoted as τ_{g_3}, with τ_{g_3} being greater than τ_{p_3} because the sensor data is considered to be more accurate. This indicates that the grid data has greater measurement uncertainty compared to the point data, which aligns with the real-world data characteristic, where sensor data are generally more accurate than satellite image data.

Table 5.2: Prior specification for the temporal coefficient in the spatio-temporal data fusion model.

Parameters	Informative prior	Non-informative prior
a_1		PC $(0.5, \alpha)$
a_2		PC $(0.5, \alpha)$
a_3		PC $(0.5, \alpha)$

Figure 5.2, 5.3 and 5.4 demonstrate how the latent fields change across space and over time. Figure 5.5 shows the spatial distribution of sparse sensor locations, and Figure 5.6 shows the coarse-resolution satellite data covering the whole study catchment. The autoregressive coefficients of each variable are $a_1 = 0.4$, $a_2 = 0.5$ and $a_3 = 0.6$ respectively, which is quite moderate temporal dependence, so the spatial pattern across different time points is not very obvious. Since the point data and grid data are a combination of all three latent fields (μ_1, μ_2, and μ_3), it is even challenging to tell the temporal trends in the point data and grid data figures.

5.3.3 Model fitting

The prior distributions for common parameters in the spatio-temporal model, such as intercepts, scaling parameters, and spatial parameters, are the same as those used in the spatio-only model described in Section 4.3.3, except for the temporal coefficients, whose priors are specified separately in Table 5.2. The penalised-complexity (PC) priors are used here for the temporal coefficients to guide the Bayesian inference process towards less complex solutions by penalising complexity and the distance from the base model by shrinking the range toward infinity and the marginal variance toward zero (Fuglstad et al., 2019) (the details of PC priors are shown in Section 3.3.4).

Table 5.3 presents the parameter estimation performance with the RMSE and mean of the spatio-temporal point, grid, and joint models across varying numbers of time points (k) on simulation datasets. Most parameter RMSE values decrease with increasing k, suggesting improved estimation accuracy. Notably, the joint model consistently outperforms both the point and grid models, with lower RMSE values in most scenarios. This performance shows the joint models ability to outperform the other two models in the spatio-temporal scenario, effectively borrowing information across both dimensions. In contrast, the point and grid models exhibit less accurate estimated parameters.

The RMSE improvement is very obvious as k increases from 3 to 10 time points, suggesting that temporal information enhances estimation by reducing uncertainty through repeated measurements on the same locations. However, beyond $k = 7$, the improvement stabilises, with little difference observed between $k = 7$ and $k = 10$. This plateau implies a threshold where additional temporal data can no longer contribute to parameter estimation, probably because the model has already captured the dominant temporal variability. From a real data application perspective, this

finding suggests that allocating resources to increase temporal resolution beyond $k = 10$ may not guarantee better model performance. Instead, refining spatial resolution or incorporating more data sources could be more helpful for further model improvements.

The joint models robustness highlights the importance of spatio-temporal frameworks in complex systems with evolving spatial patterns. The simulation results demonstrate that incorporating multiple time points not only improves predictive performance but also enhances the accuracy and stability of parameter estimation. This insight is particularly relevant for applications such as environmental monitoring or epidemiology, where reliable inference of dynamic processes is crucial.

Table 5.3: Posterior summaries (mean, 2.5% and 97.5% quantiles, posterior SD, RMSE) across different k for the spatio-temporal fusion model.

(a) Intercept parameters (α)

Parameter	True	$k=3$					$k=7$					$k=10$				
	Value	Mean	$Q_{0.025}$	$Q_{0.975}$	sd_{post}	RMSE	Mean	$Q_{0.025}$	$Q_{0.975}$	sd_{post}	RMSE	Mean	$Q_{0.025}$	$Q_{0.975}$	sd_{post}	RMSE
α_1 point	0.5	0.482	−0.395	1.343	0.443	0.421	0.450	−0.136	0.989	0.287	0.286	0.473	−0.004	0.942	0.241	0.236
α_1 grid	0.5	0.477	−0.282	1.189	0.375	0.311	0.471	−0.072	0.935	0.257	0.205	0.486	0.036	0.882	0.216	0.183
α_1 joint	0.5	0.481	−0.277	1.193	0.375	0.306	0.474	−0.096	0.966	0.271	0.207	0.491	0.011	0.900	0.227	0.186
α_2 point	0.8	0.842	0.487	1.103	0.157	0.176	0.788	0.566	1.008	0.229	0.103	0.793	0.607	1.000	0.100	0.078
α_2 grid	0.8	0.837	0.484	1.049	0.144	0.145	0.784	0.558	0.986	0.109	0.096	0.801	0.604	0.980	0.096	0.077
α_2 joint	0.8	0.839	0.516	1.020	0.129	0.146	0.790	0.546	0.977	0.110	0.094	0.799	0.569	1.015	0.114	0.072
α_3 point	1.0	1.339	0.889	2.059	0.298	0.424	1.445	1.063	1.960	0.229	0.485	1.444	1.171	1.898	0.185	0.472
α_3 grid	1.0	1.358	0.473	1.519	0.267	0.404	1.434	0.658	1.532	0.223	0.456	1.430	0.786	1.524	0.188	0.445
α_3 joint	1.0	1.384	0.685	1.619	0.238	0.427	1.446	0.906	1.676	0.196	0.473	1.442	0.995	1.650	0.167	0.457

(b) Scaling parameters (β)

Parameter	True	$k=3$					$k=7$					$k=10$				
	Value	Mean	$Q_{0.025}$	$Q_{0.975}$	sd_{post}	RMSE	Mean	$Q_{0.025}$	$Q_{0.975}$	sd_{post}	RMSE	Mean	$Q_{0.025}$	$Q_{0.975}$	sd_{post}	RMSE
β_1 point	−0.3	−0.212	−0.468	−0.078	0.099	0.138	−0.214	−0.387	−0.167	0.056	0.041	−0.180	−0.366	−0.162	0.044	0.040
β_1 grid	−0.3	−0.277	−0.382	−0.120	0.067	0.077	−0.176	−0.374	−0.212	0.041	0.028	−0.150	−0.368	−0.229	0.035	0.023
β_1 joint	−0.3	−0.279	−0.343	−0.116	0.058	0.066	−0.156	−0.338	−0.112	0.037	0.025	−0.125	−0.327	−0.105	0.031	0.016
β_2 point	−0.4	−0.308	−0.544	0.008	0.140	0.145	−0.342	−0.503	−0.160	0.087	0.080	−0.273	−0.462	−0.181	0.071	0.055
β_2 grid	−0.4	−0.327	−0.550	−0.011	0.137	0.137	−0.396	−0.557	−0.176	0.097	0.104	−0.333	−0.564	−0.236	0.083	0.080
β_2 joint	−0.4	−0.363	−0.730	−0.064	0.169	0.110	−0.415	−0.698	−0.291	0.103	0.102	−0.326	−0.662	−0.235	0.083	0.100
β_3 point	−0.2	−0.194	−0.569	0.024	0.151	0.129	−0.407	−0.485	−0.031	0.116	0.120	−0.350	−0.450	−0.079	0.095	0.069
β_3 grid	−0.2	−0.185	−0.247	0.256	0.128	0.086	−0.387	−0.201	0.218	0.107	0.077	−0.334	−0.184	0.176	0.092	0.080
β_3 joint	−0.2	−0.180	−0.254	0.188	0.113	0.089	−0.357	−0.240	0.133	0.095	0.083	−0.287	−0.226	0.096	0.078	0.055

(c) Variance parameters (σ^2)

Parameter	True	$k=3$					$k=7$					$k=10$				
	Value	Mean	$Q_{0.025}$	$Q_{0.975}$	sd_{post}	RMSE	Mean	$Q_{0.025}$	$Q_{0.975}$	sd_{post}	RMSE	Mean	$Q_{0.025}$	$Q_{0.975}$	sd_{post}	RMSE
σ_1^2 point	1.0	1.002	0.718	1.562	0.216	0.158	1.038	0.809	1.318	0.129	0.099	1.055	0.856	1.290	0.110	0.095
σ_1^2 grid	1.0	0.955	0.722	1.245	0.133	0.151	0.931	0.775	1.112	0.086	0.133	0.906	0.776	1.054	0.071	0.136
σ_1^2 joint	1.0	0.923	0.691	1.213	0.133	0.165	0.898	0.738	1.085	0.088	0.157	0.874	0.741	1.026	0.072	0.160
σ_2^2 point	0.5	0.463	0.328	0.640	0.080	0.096	0.478	0.391	0.580	0.048	0.055	0.489	0.412	0.578	0.042	0.049
σ_2^2 grid	0.5	0.451	0.323	0.614	0.074	0.071	0.447	0.363	0.545	0.046	0.070	0.444	0.371	0.525	0.039	0.064
σ_2^2 joint	0.5	0.352	0.236	0.510	0.070	0.190	0.391	0.306	0.493	0.047	0.154	0.413	0.328	0.514	0.047	0.154
σ_3^2 point	0.3	0.323	0.160	0.608	0.117	0.071	0.316	0.214	0.455	0.062	0.073	0.300	0.216	0.406	0.048	0.042
σ_3^2 grid	0.3	0.332	0.212	0.498	0.073	0.071	0.298	0.215	0.545	0.062	0.033	0.291	0.221	0.376	0.071	0.036
σ_3^2 joint	0.3	0.252	0.144	0.417	0.070	0.086	0.288	0.200	0.409	0.053	0.122	0.282	0.201	0.386	0.047	0.098

Table 5.4: Posterior summaries (mean, 2.5% and 97.5% quantiles, posterior SD, RMSE) across different k for the spatio-temporal fusion model.

(a) Range parameters (ρ)

Parameter	True	$k=3$					$k=7$					$k=10$				
	Value	Mean	$Q_{0.025}$	$Q_{0.975}$	sd_{post}	RMSE	Mean	$Q_{0.025}$	$Q_{0.975}$	sd_{post}	RMSE	Mean	$Q_{0.025}$	$Q_{0.975}$	sd_{post}	RMSE
ρ_1 point	4.0	4.908	2.246	9.590	1.913	1.744	4.309	2.760	6.425	0.936	1.143	4.055	2.817	5.646	0.721	0.809
ρ_1 grid	4.0	4.867	3.038	7.449	1.127	1.097	4.775	3.536	6.343	0.715	0.944	4.775	3.697	6.087	0.608	0.903
ρ_1 joint	4.0	5.025	3.142	7.728	1.172	1.263	5.114	3.759	6.838	0.784	1.220	5.084	3.918	6.515	0.661	1.169
ρ_2 point	3.0	3.839	1.747	7.711	1.552	1.739	3.115	2.045	4.547	0.530	0.387	3.059	2.143	4.222	0.530	0.302
ρ_2 grid	3.0	3.697	1.907	6.636	1.216	1.244	3.814	2.548	5.548	0.765	0.884	3.822	2.729	5.232	0.638	0.901
ρ_2 joint	3.0	4.574	2.141	8.928	1.760	2.520	3.830	2.452	5.755	0.843	1.420	3.889	2.692	5.471	0.708	1.201
ρ_3 point	2.0	3.305	1.139	17.847	7.064	1.733	3.002	1.626	6.268	1.212	1.480	2.603	1.586	4.561	0.763	0.866
ρ_3 grid	2.0	3.392	2.159	11.302	2.400	1.647	3.123	2.684	8.602	1.520	1.314	2.954	2.548	6.929	1.122	1.061
ρ_3 joint	2.0	2.817	1.270	18.788	4.544	1.052	2.619	2.044	8.187	1.594	0.803	2.583	1.877	6.176	1.106	0.740

(b) Temporal coefficients (a)

Parameter	True	$k=3$					$k=7$					$k=10$				
	Value	Mean	$Q_{0.025}$	$Q_{0.975}$	sd_{post}	RMSE	Mean	$Q_{0.025}$	$Q_{0.975}$	sd_{post}	RMSE	Mean	$Q_{0.025}$	$Q_{0.975}$	sd_{post}	RMSE
a_1 point	0.4	0.340	−0.153	0.734	0.232	0.274	0.331	0.050	0.577	0.135	0.197	0.343	0.097	0.558	0.118	0.171
a_1 grid	0.4	0.425	0.118	0.677	0.144	0.149	0.390	0.205	0.559	0.090	0.062	0.389	0.236	0.531	0.075	0.047
a_1 joint	0.4	0.449	0.133	0.704	0.147	0.144	0.436	0.240	0.609	0.094	0.081	0.436	0.274	0.582	0.078	0.052
a_2 point	0.5	0.402	−0.025	0.724	0.195	0.189	0.446	0.237	0.623	0.098	0.127	0.463	0.290	0.613	0.082	0.111
a_2 grid	0.5	0.419	0.044	0.705	0.171	0.260	0.438	0.195	0.646	0.115	0.112	0.456	0.258	0.627	0.094	0.067
a_2 joint	0.5	0.469	0.038	0.818	0.207	0.367	0.604	0.395	0.770	0.096	0.198	0.644	0.479	0.777	0.076	0.198
a_3 point	0.6	0.488	−0.157	0.889	0.281	0.311	0.593	0.173	0.812	0.144	0.149	0.591	0.257	0.809	0.144	0.165
a_3 grid	0.6	0.441	0.100	0.889	0.211	0.401	0.574	0.534	0.924	0.102	0.146	0.594	0.568	0.900	0.086	0.139
a_3 joint	0.6	0.585	0.141	0.969	0.236	0.205	0.610	0.560	0.960	0.057	0.098	0.618	0.733	0.948	0.056	0.073

Note: $Q_{0.025}$ and $Q_{0.975}$ are the 2.5% and 97.5% posterior quantiles; sd_{post} is the posterior standard deviation; RMSE combines bias and variance.

5.3.3.1 Scenario 1: Final-day predictions at unobserved locations on the last day of the training period with varying time points ($k = 3, 7, 10, 30$).

The first step to validate the spatio-temporal data fusion model is to assess its predictive performance at the time boundary of the training period. In Scenario 1, the whole dataset spans 100 days, with the final day used as the test set. From this day, 20 unobserved locations are randomly selected as test points. The training set consists of the last k days, including the final day, and is used to train the model. Predictions are made for the test points on the last day. The INLA-SPDE framework integrates spatially structured random effects via the SPDE approach, which approximates the Matérn covariance field using Gaussian Markov random fields (GMRFs), while temporal dependencies are modelled through an autoregressive process (AR(1)). Predictions at the test set are generated by sampling from the posterior distribution of the latent field. The root mean squared prediction error (RMSPE) is used to quantify performance.

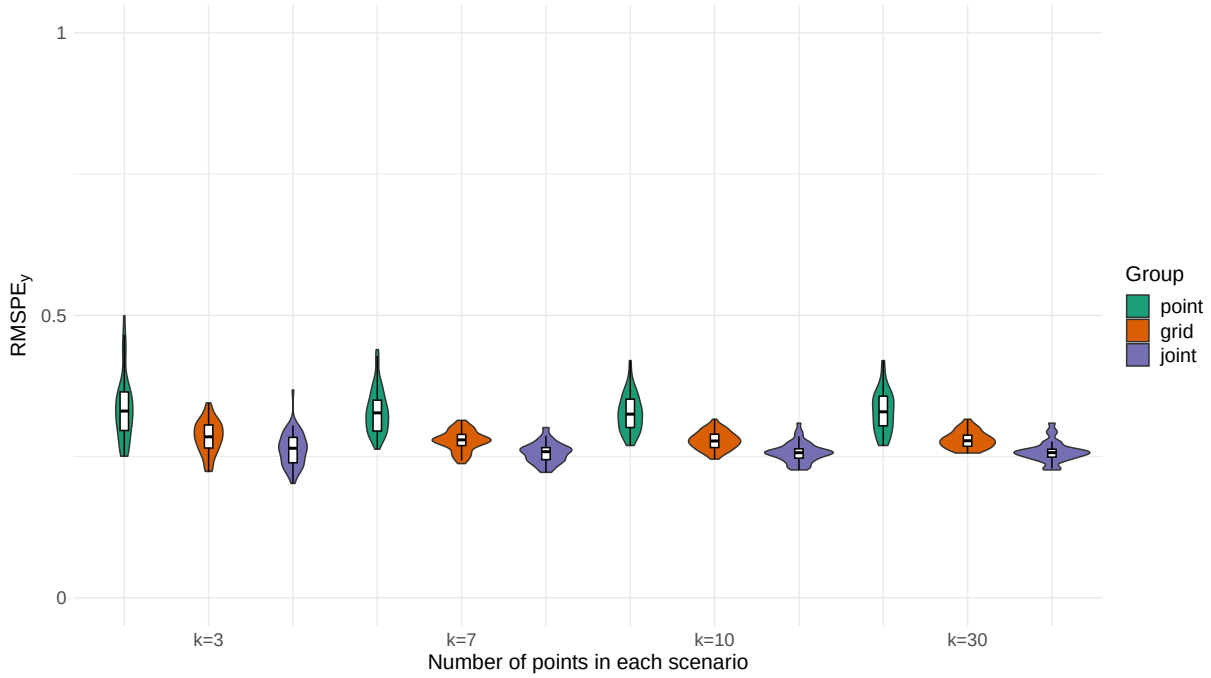


Figure 5.7: Comparison of prediction error (RMSPE$_y$) at unobserved locations of point, grid, and joint models with varying numbers of time points (k=3, 7, 10, 30). Results based on 100 simulations with medium variance latent field ($\sigma_1 = 1, \sigma_2 = 0.5, \sigma_3 = 0.3$) of final day predictions of the last day of the training period.

Figure 5.7 compares the prediction accuracy of point, grid, and joint models when predicting values at the unobserved locations of the last day of the training period. The violin plots show prediction errors (RMSPE$_y$) across 100 simulations, with lower values indicating better performance. As the number of time points increases (from $k = 3$ to $k = 30$), the joint model consistently outperforms the others, while the point model shows little improvement. As the number of time points increases, the height of the joint models violin plot decreases, indicating less variability in the simulations. This demonstrates that the joint model is more effective at using temporal information to reduce uncertainty when compared to the point and grid models.

5.3.3.2 Scenario 2: One-day-ahead future predictions at unobserved locations beyond the training period with varying time points ($k = 3, 7, 10, 30$).

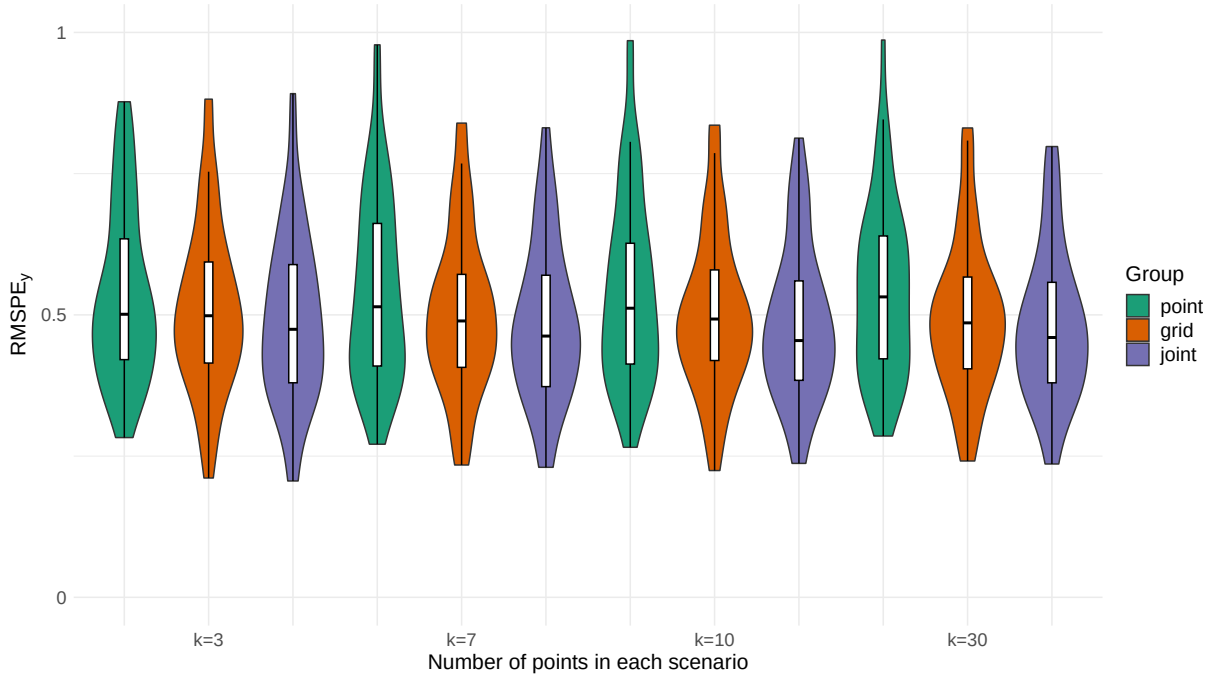


Figure 5.8: Comparison of prediction error (RMSPE$_y$) at unobserved locations of point, grid, and joint models with varying numbers of time points (k=3, 7, 10, 30). Results based on 100 simulations with medium variance latent field ($\sigma_1 = 1, \sigma_2 = 0.5, \sigma_3 = 0.3$) of one-day-ahead predictions of the training period.

In Scenario 2, the dataset covers 100 days, with the final day used as the test set. On this day, 20 unobserved locations are randomly selected as test points. The training set consists of the k days immediately before the final day (it excludes the last day) and is used to fit the model. Predictions are then made for the test points on the final day.

Figure 5.8 shows the one-day-ahead prediction errors (RMSPE$_y$) at unobserved locations. The joint model still outperforms or performs equally well as the point and grid model, though with smaller margins than in Figure 5.7. All models show higher overall error (0.4-0.6 range) for future predictions beyond the training period, with performance differences remaining relatively stable despite increasing time points. The distributions of the RMSPE$_y$ appear wider, suggesting more variability in one-day-ahead prediction accuracy. This suggests that while temporal information improves model performance, forecasting future time points remains more challenging and uncertain.

Table 5.5: Joint model parameter estimates: Gridded covariates missing vs. Gridded covariates complete (k = 3)

		Gridded covariates missing					Gridded covariates complete				
Parameter	True	Mean	$Q_{0.025}$	$Q_{0.975}$	SD_{post}	RMSE	Mean	$Q_{0.025}$	$Q_{0.975}$	SD_{post}	RMSE
α_1 joint	0.50	0.43738	−0.41902	1.29379	0.43695	0.35670	0.481	−0.10197	1.24067	0.34252	0.306
α_2 joint	0.80	0.78500	0.47241	1.09759	0.15949	0.14594	0.839	0.60977	1.02214	0.10520	0.146
α_3 joint	1.00	0.98299	0.40275	1.56325	0.29605	0.23692	1.384	0.685	1.619	0.238	0.427
β_1 joint	−0.30	−0.16089	−0.33743	0.01636	0.08986	0.17042	−0.279	−0.343	−0.116	0.058	0.066
β_2 joint	−0.40	−0.22929	−0.51917	0.06246	0.14772	0.25781	−0.363	−0.730	−0.064	0.169	0.110
β_3 joint	−0.20	−0.02157	−0.29190	0.24876	0.13793	0.20583	−0.180	−0.254	0.188	0.113	0.089
σ_1^2 joint	1.00	0.97074	0.61974	1.47579	0.21916	0.19287	0.923	0.691	1.213	0.133	0.165
σ_2^2 joint	0.50	0.46470	0.31941	0.66096	0.08713	0.09414	0.352	0.236	0.510	0.070	0.190
σ_3^2 joint	0.30	0.36616	0.22047	0.58274	0.09289	0.11837	0.252	0.144	0.417	0.070	0.086
ρ_1 joint	4.00	5.55047	2.46004	10.91789	2.18729	2.12348	5.025	3.142	7.728	1.172	1.263
ρ_2 joint	3.00	3.71899	1.68613	7.41643	1.49432	1.67957	4.574	2.141	8.928	1.760	2.520
ρ_3 joint	2.00	5.38398	2.00989	12.75632	2.87019	5.52838	2.817	1.270	18.788	4.544	1.052
a_1 joint	0.40	0.32715	−0.24163	0.78535	0.27351	0.25807	0.449	0.133	0.704	0.147	0.144
a_2 joint	0.50	0.44840	0.02347	0.74972	0.18929	0.23608	0.469	0.038	0.818	0.207	0.367
a_3 joint	0.60	0.75636	0.35093	0.94594	0.15742	0.19790	0.585	0.141	0.969	0.236	0.205

5.3.3.3 Assessing the joint model performance under realistic grid covariate missingness

In the real data application, the only available satellite data is the response variable Soil Water Index (SWI), while all the covariates (e.g., rainfall and temperature) from satellite data are not available. To ensure the proposed model framework remains robust under this setting, a targeted simulation study was designed to mimic the real data scenario. Specifically, we compare the performance of the joint model in two situations: one where the grid covariates are available, and another where they are completely missing. The point level data, including both responses and covariates, are kept identical across both situations.

To reflect the structure of the real application, we fix the number of time points at $k = 3$. The only difference between the two scenarios is the availability of the grid covariates. Table 5.5 presents the posterior summaries of the joint model parameters under two scenarios: with and without the grid covariates. Each parameter estimate is listed along with its posterior mean, standard deviation, 95% credible interval (defined by the 2.5% and 97.5% quantiles), and root mean squared error (RMSE).

The results show that the joint model produces reasonably accurate estimates for the intercept parameters α_1, α_2, and α_3 under both scenarios. However, under the grid missing condition, posterior uncertainty increases slightly, and RMSEs are higher, particularly for α_3, which appears more sensitive to missing grid-level covariates.

In contrast, the scaling parameters β_1, β_2, and β_3 are more affected by missing grid covariates. These parameters show greater posterior variability and higher RMSEs in the grid missing situation, indicating reduced identifiability when covariate information is incomplete.

The spatial variance parameters σ_1^2, σ_2^2, and σ_3^2 are estimated reasonably well in both scenarios. Although credible intervals are slightly wider and RMSEs slightly higher under the grid missing situation, the estimates remain close to the true values, suggesting the model maintains robustness for these latent fields' variance terms.

However, the spatial range parameters ρ_1, ρ_2, and ρ_3 are poorly recovered when the gridded covariates are missing, with posterior means overestimating the true values and large RMSEs. This suggests that the range parameters are structurally difficult to identify in this setting, likely due to the fact that when gridded covariates are missing, the latent fields lose large-scale structure, which might produce overestimation of ρ.

Finally, the temporal coefficients a_1, a_2, and a_3 remain relatively stable between the two scenarios. The posteriors and RMSEs change only slightly when gridded covariates are added in the joint model, suggesting inference in the joint model is dominated by the point data.

In summary, the comparison suggests that the joint model is robust to missing grid-level covariates and produces stable inferences even when covariate information is partially missing.

5.4 Real data application

In Chapter 4, the real data application is conducted using the spatio-only model, using soil moisture data from a single day to investigate spatial dependencies across the whole study catchment. This spatio-only data fusion model provides valuable insights into the models ability to capture spatial variation, but does not account for temporal dynamics. In this section, the real data application is extended by incorporating multiple days of soil moisture data, allowing for the evaluation of temporal information in prediction modelling. By including data from multiple time points, the spatio-temporal data fusion model aims to determine whether modelling temporal dependencies alongside spatial correlation leads to better predictive performance on the real datasets. Specifically, the temporal information enables the model to potentially leverage patterns such as soil moisture persistence, seasonal effects, or delayed responses to covariates (rainfall). Through this comparison between the spatio-only and spatio-temporal models, this section aims to quantify the gains in prediction accuracy from the incorporation of temporal structure, thereby providing deep insights into the advantages of spatio-temporal modelling for soil moisture data fusion.

Figure 5.9 presents the one-day-ahead prediction map based on a 10-day training set from 06/05/2022 to 15/05/2022. It is noted that the grid covariates of the satellite data are missing, so the only available information of the grid data is the response variable SWI. The left column

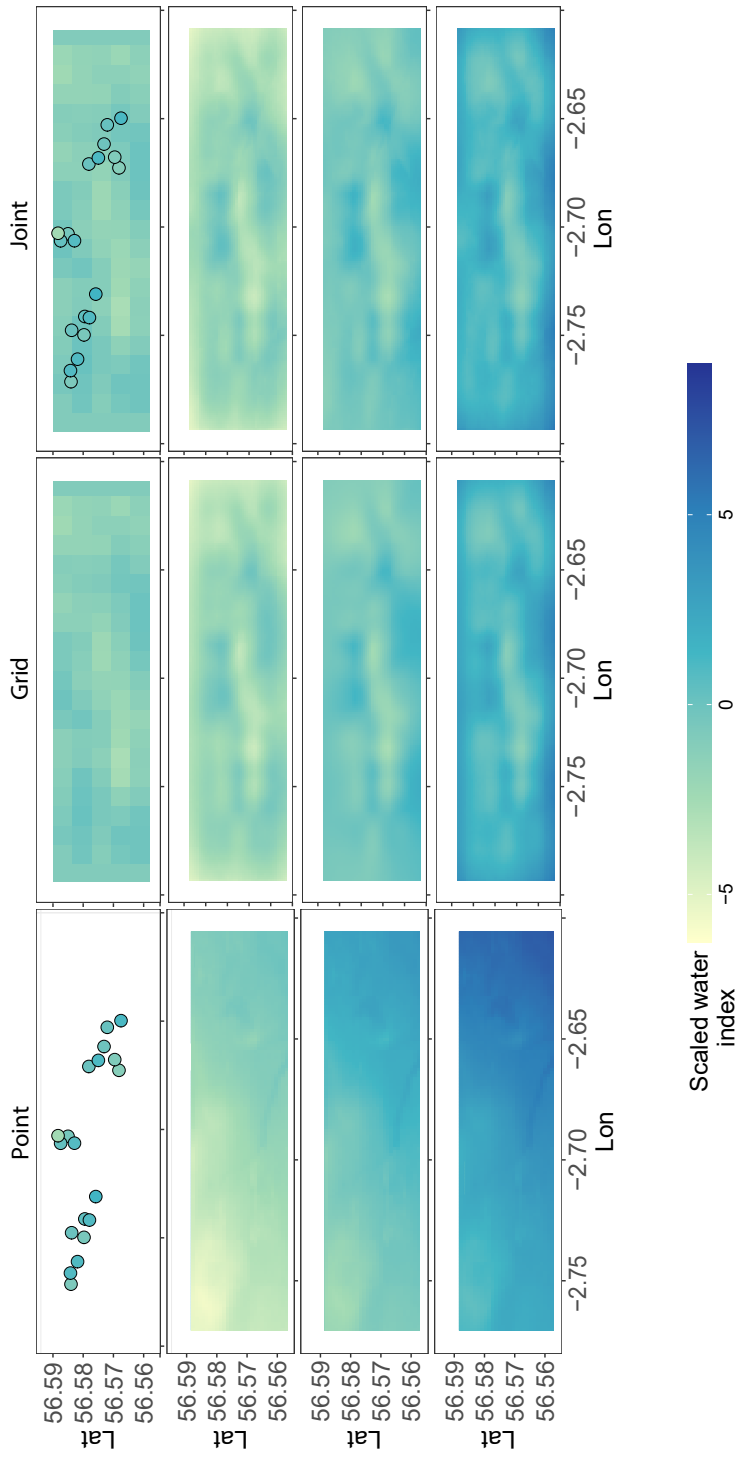


Figure 5.9: Prediction maps of the standardised water index (VWC and SWI) over the Elliott water catchment area on 16/06/2022 (the 10-day training set includes data from 06/05/2022 to 15/06/2022), using the same mesh, with 95% credible intervals. The top row displays the real data for each model, while the subsequent rows show the 95% prediction interval, including the 2.5% and 97.5% quantiles as well as the mean of the predictions.

displays the raw observations, the 95% prediction interval, and the mean predictions for the point model, which reveals that the spatial pattern is strongly dominated by elevation. The grid model captures more spatial detail due to its broader spatial coverage. The joint model, which integrates both point and grid data, narrows the CIs of the prediction by using the strengths of both datasets. Table 5.6 shows the posterior means along with the 95% credible intervals (2.5% and 97.5% quantiles) for the spatio-temporal model with different time points ($k = 3$ and $k = 10$). The parameters are grouped into intercepts, scaling parameters, spatial parameters, and temporal coefficients. Each parameter is evaluated across three models (joint, point, and grid) within the spatio-temporal model framework, with the model fitted for time points of $k = 3$ and $k = 10$. The intercept estimates across the point grid and the joint model vary, and for most of the intercepts, there is less uncertainty with $k = 10$ than with $k = 3$. The scaling parameters suggest that the effects are different between $k = 3$ and $k = 10$, with the exception that β_3 remains negative across both values of k, suggesting a consistent negative effect. The range (ρ) and variance (σ) also vary a lot, and the temporal coefficients (a_3) indicate strong temporal autocorrelation for y_3.

The real data application reveals a trade-off between temporal autocorrelation and parameter uncertainty in the spatio-temporal model. While increasing the number of time points ($k = 10$) improves precision for parameters with strong temporal persistence (e.g., coefficients like a_3), it increases uncertainty for intercepts (α_3). The conflict reflects model structural constraints introduced by more time points (sparse daily point data, weak identifiability or stronger spatio-temporal interactions). In addition, the spatial range may change over time, yet the model assumes it is fixed. For weakly identified parameters (e.g., β_1 and β_2), additional time points lead to unstable parameter estimation, while strongly autocorrelated temporal processes ($a_3 \approx 1$) benefited from more time points. The results highlight the importance of balancing model complexity with data adequacy: more temporal data improves signals for dominant processes but increases noise in hierarchical parameters, which needs careful prior specification or model redesign to stabilise inferences.

Overall, increasing the time points does not guarantee decreasing the uncertainty for all parameters. Adding more time points provides more repeated measures over time, which contributes to the estimation of the temporal coefficients by capturing long-term patterns and reduces the uncertainty of the parameter estimation (e.g., narrower CIs for temporal coefficients a). In Table 5.6, the a_1 of the joint model at $k = 10$ has a posterior mean with a tight interval compared to $k = 3$, where the CIs are wider. For example, when $k = 3$, the temporal coefficient a_1 from the joint model is estimated at 0.53 with a 95% CI of (-0.46, 0.99), whereas for $k = 10$, it increases to 0.98 with a narrower CI of (0.94, 0.99). This indicates that the certainty in the temporal coefficients improves as the number of time points increases in the model. In Table 5.6c, the spatial parameters do not show decreasing uncertainty with the number of time points increasing, which could be caused by introducing more parameters that may not be fully identifiable.

Table 5.6: Parameter estimates with posterior means and 95% credible intervals, obtained by fitting the point, grid and joint spatio-temporal data fusion models to the soil moisture data.

(a) Intercepts

Parameter	$k = 3$	$k = 10$
α_1 joint	$-0.55\,(-0.59, -0.51)$	$-0.30\,(-0.63, 0.02)$
α_1 point	$-0.55\,(-0.59, -0.51)$	$-0.29\,(-0.42, -0.16)$
α_1 grid	–	–
α_2 joint	$0.18\,(-1.07, 1.42)$	$0.12\,(-0.17, 0.42)$
α_2 point	$0.12\,(-0.34, 0.58)$	$0.02\,(-0.18, 0.22)$
α_2 grid	–	–
α_3 joint	$0.29\,(-4.18, 4.76)$	$0.17\,(-7.88, 8.22)$
α_3 point	$0.33\,(-0.92, 1.58)$	$0.33\,(-4.30, 4.96)$
α_3 grid	$0.42\,(-3.90, 4.75)$	$0.22\,(-7.53, 7.98)$

(b) Scaling parameters

Parameter	$k = 3$	$k = 10$
β_1 joint	$-0.05\,(-0.50, 0.39)$	$2.51\,(2.18, 2.83)$
β_1 point	$0.01\,(-0.55, 0.58)$	$0.11\,(-0.40, 0.66)$
β_2 joint	$-0.81\,(-1.28, -0.33)$	$0.60\,(0.38, 0.85)$
β_2 point	$-0.07\,(-0.35, 0.22)$	$0.27\,(0.15, 0.38)$
β_3 joint	$-0.85\,(-1.06, -0.63)$	$-0.35\,(-0.52, -0.17)$
β_3 point	$-1.70\,(-2.18, -1.22)$	$-1.18\,(-1.38, -0.98)$
β_3 grid	$-0.14\,(-0.69, 0.42)$	$-0.14\,(-0.52, 0.24)$

(c) Spatial parameters (variance and range)

Parameter	$k = 3$	$k = 10$
σ_1^2 joint	$0.03\,(0.01, 0.07)$	$1.31\,(0.78, 2.07)$
σ_1^2 point	$0.02\,(0.01, 0.05)$	$0.11\,(0.05, 0.20)$
σ_2^2 joint	$0.92\,(0.41, 1.82)$	$0.98\,(0.80, 1.18)$
σ_2^2 point	$0.92\,(0.61, 1.30)$	$0.93\,(0.80, 1.09)$
σ_3^2 joint	$2.89\,(2.09, 3.91)$	$3.13\,(2.47, 3.92)$
σ_3^2 point	$1.13\,(0.73, 1.69)$	$0.33\,(0.12, 0.67)$
σ_3^2 grid	$2.84\,(2.05, 3.87)$	$3.09\,(2.44, 3.87)$
ρ_1 joint	$4.2 \times 10^4\,(1.3 \times 10^4, 1.2 \times 10^5)$	$2.1 \times 10^4\,(1.5 \times 10^4, 2.8 \times 10^4)$
ρ_1 point	$3.6 \times 10^4\,(1.1 \times 10^4, 9.8 \times 10^4)$	$4.3 \times 10^4\,(2.0 \times 10^4, 9.1 \times 10^4)$
ρ_1 grid	–	–
ρ_2 joint	$8.1 \times 10^3\,(2.9 \times 10^3, 1.9 \times 10^4)$	$3.5 \times 10^3\,(2.6 \times 10^3, 4.7 \times 10^3)$
ρ_2 point	$1.8 \times 10^3\,(6.0 \times 10^2, 4.3 \times 10^3)$	$9.2 \times 10^2\,(6.1 \times 10^2, 1.3 \times 10^3)$
ρ_2 grid	–	–
ρ_3 point	$7.8 \times 10^3\,(3.5 \times 10^3, 1.5 \times 10^4)$	$8.3 \times 10^4\,(3.7 \times 10^4, 1.8 \times 10^5)$
ρ_3 joint	$6.2 \times 10^3\,(4.6 \times 10^3, 8.4 \times 10^3)$	$1.1 \times 10^4\,(8.7 \times 10^3, 1.3 \times 10^4)$
ρ_3 grid	–	–

Table 5.6: Parameter estimates with posterior means and 95% credible intervals (2.5% and 97.5% quantiles), obtained by fitting the point, grid and joint spatio-temporal data-fusion models to the soil moisture data. (continued)

(d) Temporal coefficient parameters

Parameter	k=3	k=10
a_1 joint	0.53 (-0.46, 0.99)	0.98 (0.94, 0.99)
a_1 point	0.36 (-0.51, 0.92)	0.33 (-0.20, 0.78)
a_1 grid	-	-
a_2 joint	0.46 (-0.43, 0.94)	-0.22 (-0.51, 0.07)
a_2 point	0.09 (-0.33, 0.48)	0.07 (-0.09, 0.23)
a_2 grid	-	-
a_3 joint	0.98 (0.97, 0.99)	0.98 (0.97, 0.99)
a_3 point	0.05 (-0.51, 0.60)	0.92 (0.73, 0.99)
a_3 grid	0.98 (0.97, 0.99)	0.98 (0.97, 0.99)

5.5 Conclusion

This chapter develops a spatio-temporal data fusion framework that builds directly on the spatial-only model of Chapter 4. Key features in the model include latent Gaussian random fields with Matérn spatial covariances extended over time via an AR(1) structure, and a fusion strategy that integrates point-level and gridded remote-sensing observations. We assume separable spatio-temporal dependence. To be specific, independent AR(1) evolves in time along with a stationary Matérn covariance structure at each time point. Limitations of the current approach include the fixed smoothness parameter in the Matérn covariance. After introducing the models (Section 5.2), we assess model predictive performance through a simulation study and real-data application.

The simulation study systematically evaluated the performance of point, grid, and joint models across different numbers of time points ($k = 3, 7, 10, 30$), focusing on the model ability to estimate parameters within the model, including intercepts ($\alpha_1, \alpha_2, \alpha_3$), scaling parameters ($\beta_1, \beta_2$), spatial variances ($\sigma^2$), and range parameters (ρ). The results show that increasing the number of time points generally improves parameter estimation accuracy, with root mean squared error (RMSE) and bias reducing very notably for parameters such as α_1, α_2, β_1, and β_2. For example, the RMSE for β_1 decreases by around 35% when increasing from $k = 3$ to $k = 30$, highlighting the worth of including temporal information for true parameter recovery and prediction accuracy. However, not all parameters benefit from the increasing number of time points: biases in α_3 and ρ are still there regardless of the number of time points, suggesting that these parameters may be more sensitive to model assumptions or structural constraints rather than the amount of temporal information available.

Among the point model, grid model and joint model, the joint model consistently shows the best

performance. At $k = 30$, it achieves RMSE decreasing by 15%–20% for scaling parameters and 10%–15% for spatial variances compared to the point and grid models. The model benefits from its ability to integrate spatial and temporal processes through a shared latent structure, making it more robust when the data are sparse. In contrast, the grid and point models show greater sensitivity to limited time points, with the grid model's RMSE for ρ larger than that of the joint model by roughly 25% at $k = 3$.

Adding more time points narrows posterior credible intervals for most of the parameters. For example, the 95% CIs for α_1 decreases from $[-0.59, -0.51]$ at $k = 3$ to $[-0.27, -0.16]$ at $k = 30$. However, it does not solve the structural biases. Errors in parameters like α_3 and ρ indicate that simply increasing the number of time points cannot fully compensate for limitations in model design, such as oversimplified temporal covariance functions or weakly informative priors.

There are several directions worth exploring. Model structure improvement, such as incorporating spatially adaptive range parameters or higher-order temporal dependencies, may help address biases in ρ and α_3. Moreover, testing the model under non-stationary conditions and irregular sampling strategies will be important to evaluate model generalisability. Finally, comparing this model with benchmark deep learning architectures, such as spatio-temporal Transformers, could offer valuable insights into balancing interpretability and predictive performance.

In summary, while increasing time points improves the predictive performance of the spatio-temporal data fusion model, it also suggests model structure limitations that require methodology innovations. The joint model outperforms the point model and grid model and can be regarded as a robust choice for spatio-temporal data fusion. However, the residual biases suggest unresolved challenges. Future work should focus on improving both computational scalability and model flexibility to better support complex real-world applications.

Chapter 6

Spatio-temporally constrained ensemble learning with conformal prediction: A distribution-free approach to uncertainty-aware data fusion

6.1 Introduction

Spatial misalignment is a challenge when fusing datasets with different spatial supports, and the existing literature gives several methods to address this issue. Traditional statistical methods, such as Kriging (Stein, 1999), can do interpolation for point data and accommodate support differences through adaptations such as block kriging. However, these methods require solving a Kriging system for each data point, which leads to high computational costs. Additionally, Kriging is limited by its strict assumptions and lack of flexibility, which makes it difficult to capture complex, nonlinear relationships and interactions between variables.

Another model is the Bayesian hierarchical models (BHM), which provide a flexible framework by incorporating latent spatial processes and explicitly quantifying uncertainty. This model structure can do seamless integration of data collected at different spatial scales through a three-layer framework, which includes data, latent process, and parameter models. However, these models can be computationally expensive, particularly when applied to large-scale datasets.

Recently, modern machine learning, such as neural networks and XGBoost, has overcome many of these challenges with robust performance and efficiency (Chen, 2016). In this work, we will focus on XGBoost because it performs well with limited data points (regularised trees) and is fast with large-scale data (parallel boosting), which suits our real data application. It is suitable for large-scale datasets, which plays the trade-off between optimised gradient boosting

and parallel processing for fast convergence. The built-in regularisation helps reduce overfitting, and it effectively handles missing data while integrating with multiple data sources. However, its application to spatially misaligned data is limited because it does not inherently incorporate spatial dependence structures or provide uncertainty quantification.

However, conformal prediction offers a robust framework for uncertainty quantification by providing prediction intervals which capture the true outcomes with a predefined probability level (Shafer and Vovk, 2008). This approach is grounded in rigorous statistical theory, which ensures that the estimated intervals have valid coverage properties even with minimal distributional assumptions (Mao et al., 2024).

Chapter 3 introduces a framework to address spatial misalignment in spatial regression. Chapters 4 and 5 extend this to spatial-only and spatio-temporal data fusion models for point and gridded data. Although these frameworks perform well in simulations and real applications, their computational cost is high. Therefore, this chapter proposes the development of hybrid frameworks that integrate spatial-temporal information to bridge the gap between modern machine learning and established statistical methodologies. We also propose a spatio-temporal conformal inference to quantify the uncertainty. At the end of this chapter, we compare the predictive performance of a modern machine learning approach (XGBoost with conformal prediction) against the established BHM model (in Chapter 5).

6.2 Literature review

This section is a focused literature review limited to methods used in this chapter: XGBoost for spatio-temporal data and conformal prediction for model uncertainty. The literature review of the BHM and alternative ML methods is in the previous chapter.

6.2.1 XGBoost

Tree-based models have a long history in machine learning due to their interpretability and adaptability. The modern gradient-boosting frameworks, such as extreme gradient boosting (XGBoost), can be traced back to early decision trees. For example, Amedeo and Golledge (1975) split data using features that maximise the gain in information. Still, this method has three main disadvantages: It is easy to overfit due to its sensitivity to the training data, and it is not stable because small data changes can lead to very different trees. It has limited prediction power because a single tree fails to capture very complex patterns. To reduce the variance, Breiman (1996) perform prediction using bootstrapped datasets. Breiman (2001) extend this by randomly selecting features during splits to improve the model's robustness. These methods reduce the overfitting problem and make parallel training possible, but they lose interpretability and are

computationally expensive for large datasets. Friedman (2001) develop gradient boosting as an additive model trained using gradient descent in function spaces. The model builds an ensemble of M trees to minimise a differentiable loss function $\mathscr{L}$ iteratively.

But there are some limitations, such as being computationally expensive due to greedy split search, no explicit regularisation (which will lead to the overfitting problem), and only relying on first-order gradients. This method increases flexibility by using diverse loss functions and improves accuracy by utilising multiple and deep trees to capture nonlinear relationships. The Greedy tree-building makes it too slow for the large datasets, and there is no built-in regularisation between the depth of the tree, and it performs poorly on the missing values.

Chen (2016) make critical improvements to GBM's framework, which includes the regularised objective function, second-order Taylor approximation, and approximate split-finding with gain maximisation. The details of each improvement are as follows:

1. Regularisation: Explicit control of model complexity via γ and λ.
2. Second-order optimisation: Faster convergence using Hessian-aware updates.
3. Efficient splitting: Approximate algorithms reduce computation from $\mathscr{O}(n)$ to $\mathscr{O}(\sqrt{n})$ per split.

XGBoost has a great performance in modelling structured data and non-linear relationships, which makes it a popular method for many predictive tasks. Existing studies show XGBoost's ability to integrate the spatial-temporal features (e.g., lagged variables, geographic coordinates) to model spatial-temporal dependencies that traditional statistical methods struggle to capture. For example, studies in environmental monitoring and urban planning have successfully combined XGBoost with spatial interpolation techniques (e.g., kriging) to enhance prediction accuracy (Wong et al., 2021; Wang et al., 2023). However, a key limitation of XGBoost is that it lacks an inherent spatial-temporal structure to capture the spatial dependence (such as the adjacent matrix). While XGBoost has advantages in interoperability and scalability, existing studies point out that XGBoost is insufficient to capture spatial-temporal patterns (Meyer and Pebesma, 2022; Jemeljanova et al., 2024). To address this, a hybrid modelling framework is needed, such as combining XGBoost with domain-specific spatial-temporal dependencies. Dai et al. (2023) implement the XGBoost on multiple data sources for prediction, but they only use the point sensor data. Some other studies use grid satellite data without considering the nature of the grid data (Shetty et al., 2024).

6.2.2 Conformal prediction

Conformal prediction (CP) is a distribution-free, model-agnostic framework for generating statistical prediction intervals with coverage guarantees. Based on algorithmic learning theory, it

bridges frequentist statistics and machine learning, offering an approach to uncertainty quantification. This review traces its theoretical foundations, key advancements, and practical applications. Conformal prediction can be traced back to the work of Vovk et al. (2005), which is inspired by transductive inference and online learning principles. The framework was formalised in the context of confidence machines. Let $\mathscr{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^{m}$ be the held-out calibration set, and define the calibration scores $S_i = s(x_i, y_i)$ (e.g., $S_i = |y_i - \hat{f}(x_i)|$, or $S_i = |y_i - \hat{f}(x_i)|/\hat{\sigma}(x_i)$ for a studentised score), $i = 1, \ldots, m$. CP uses a nonconformity measure (e.g., prediction residuals) to quantify how strange a new example is relative to a calibration set. It then constructs prediction intervals by thresholding these scores to obtain a target coverage $1 - \alpha$. Given calibration scores $S_i = s(x_i, y_i)$ (e.g., $s(x,y) = |y - \hat{f}(x)|$) for $i = 1, \ldots, m$, set $\hat{q}_{1-\alpha} = \text{Quantile}_{\lceil (m+1)(1-\alpha) \rceil / (m+1)}\{S_1, \ldots, S_m\}$ and define $\mathscr{C}_\alpha(x) = \{y : \ s(x,y) \leq \hat{q}_{1-\alpha}\}$ (e.g., $\alpha = 0.05 \Rightarrow 95\%$ coverage). The key theoretical guarantee is the marginal coverage and exchangeability assumption. CP assumes data points are exchangeable (a weaker condition than i.i.d.), making it robust to many real-world scenarios.

The key theoretical guarantee is the marginal coverage: For exchangeable data, CP guarantees that the prediction interval contains the true label with probability $1 - \alpha$:

$$\mathbb{P}\left(Y_{\text{test}} \in C\left(X_{\text{test}}\right)\right) \geq 1 - \alpha,$$

where α is the significance level and C is the conformal prediction interval which maps an input X_{test} to a set of outputs.

Compared to the traditional methods, CP has two advantages. Firstly, CP has a distribution-free marginal coverage guarantee, unlike the Bayesian methods, which require likelihood specifications. Secondly, CP is model agnostic, which means it is compatible with complex model frameworks such as support vector machines (SVMs) and neural networks without asymptotic approximations. However, it comes with some requirements: it requires the exchangeability of the data, and the guaranteed coverage may sacrifice the width of the intervals. Additionally, full CP is expensive for large datasets due to the recalibration across observations in the test set.

A recent study (Lou et al., 2024b) introduces distance-decaying geographic weights to CP to relax strict exchangeability through localised calibration, which prioritises nearby calibration data points (aligned with Tobler's First Law of geography). The method bridges traditional CP's theoretical guarantees with the geospatial data structure. The key adjustment of this paper is the weighted quantiles, which are defined as follows:

$$\text{GeoQuantile}_{1-\varepsilon}\left(u_{\text{test}}, v_{\text{test}}\right) = \text{Quantile}_{1-\varepsilon}\left(\sum_{i=1}^{m} w_i\left(u_{\text{test}}, v_{\text{test}}\right) \cdot \delta_{\alpha_i}\right)$$

The main idea is that prediction intervals are computed using a geographically weighted empirical distribution of nonconformity scores. Additionally, the uncertainty varies spatially, which

reflects local data density and spatial dependence (e.g., higher uncertainty in regions with sparse calibration data).

6.3 Methodology

XGBoost, a tree-based ensemble learning algorithm, is selected for its robust performance in tabular data regression tasks. It can model nonlinear relationships and has the flexibility to accommodate heterogeneous features. In addition, XGBoost supports the use of custom objective functions, which enables the integration of domain-specific constraints such as spatial smoothness.

The aim of this approach is to generate spatially continuous soil moisture predictions with uncertainty quantification. To achieve this, we combine XGBoost with a spatial penalty loss to capture spatio-temporal dependence and conformal prediction to provide reliable uncertainty quantification. This section lists all methods used in the chapter: an adapted XGBoost with a custom loss function for spatial smoothness, K-nearest-neighbours to turn sensor point values into gridded values by averaging the k nearest sensors to each cell, and conformal prediction for uncertainty quantification.

6.3.1 Geo XGBoost

To account for spatio-temporal dependence in the data, we construct an adapted XGBoost. The original XGBoost setup will be introduced first, and then a custom loss will be detailed.

The XGBoost model forms an ensemble prediction as:

$$\boldsymbol{y} = \phi(\boldsymbol{x}) = \sum_{k=1}^{K} f_k(\mathbf{x}), \quad f^{(k)} \in \mathscr{F}, \tag{6.1}$$

where f_k is a tree parameterised by its structure and leaf weights, and prediction is the sum of the leaf weights reached across all K trees.

XGBoost minimises an objective function that is the sum of a loss term and a regularisation term:

$$\mathscr{L}(\phi) = L_S(\mathbf{y},\hat{\mathbf{y}}) + \sum_{k=1}^{K} \Omega(f_k),$$

$$\Omega(f) = \gamma J + \frac{\lambda_{\text{xgb}}}{2} \sum_{j=1}^{J} w_{\text{leaf}_j}^2,$$

where J is the number of leaves in the tree, w_{leaf_j} is the leaf weight, and γ and λ_{xgb} are regularisation parameters.

At iteration t, the prediction is updated as follows:

$$\hat{\mathbf{y}}^{(t)} = \hat{\mathbf{y}}^{(t-1)} + f_t(\mathbf{x})$$

Using a second-order Taylor expansion, the approximate objective function at iteration t is:

$$\mathscr{L}^{(t)} \approx [l(\boldsymbol{y}, \hat{\boldsymbol{y}}^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t(\mathbf{x}_i)^2] + \Omega(f_t)$$

where $g_i = \frac{\partial \ell(y_i,\hat{y})}{\partial \hat{y}}\Big|_{\hat{y}=\hat{y}_i^{(t-1)}}$ and $h_i = \frac{\partial^2 \ell(y_i,\hat{y})}{\partial \hat{y}^2}\Big|_{\hat{y}=\hat{y}_i^{(t-1)}}$.

6.3.1.1 XGBoost with customised loss function

In the default form of the XGBoost in Eq.(6.1), XGBoost minimises the following regularised loss:

$$\mathscr{L}_{\text{standard}}(\hat{y}, y) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{k=1}^{K} \Omega(f_k) \tag{6.2}$$

where $\hat{y}_i$ is the prediction for sample i, y_i is the true value, and $\Omega(f_k)$ is a complexity penalty on each base learner f_k. However, it does not account for the spatial structure of the prediction domain.

To be specific, the total objective at iteration t is:

$$\mathscr{L}^{(t)} = \sum_{i=1}^{n} \underbrace{l\big(y_i,\, \hat{y}_i^{(t-1)} + f_t(x_i)\big)}_{\text{each sample } i \text{ loss}} + \sum_{k=1}^{t} \Omega\big(f_k\big)\,.$$

The XGBoost computes g_i and h_i for each i in isolation, then build trees using sums

$$G_\ell = \sum_{i \in \text{leaf}\ell} g_i, \quad H_\ell = \sum_{i \in \text{leaf}\ell} h_i,$$

when calculating the optimal weight for leaf ℓ.

Because the loss function is applied independently to each sample i, the partial derivatives $\frac{\partial^2 \mathscr{L}}{\partial f(x_i) \partial f(x_j)}$ are 0 when $i \neq j$. Thus, the Hessian matrix with respect to the prediction vector is diagonal. XGBoost only uses those second derivatives.

$$h_i = \frac{\partial^2 l\,(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}\Big|_{\hat{y}_i=\hat{y}_i^{(t-1)}}$$

to decide how to split and assign weights to each tree.

The non-diagonal Hessian would denote interactions between samples. For example, the off-diagonal entries $H_{ij} \neq 0$ denote the curvature of the loss for sample i that depends on sample j's prediction. That breaks the assumption on which XGBoost is built (loss separability).

Our new contribution is the addition of a graph-Laplacian smoothing term to the XGBoost loss function. This penalises rapid changes between predictions at neighbouring locations, yielding spatially smoother maps. A common choice is a quadratic (Laplacian-type) penalty (Shi and Malik, 2000):

$$\text{Spatial-penalty} = \frac{\lambda}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left(\hat{y}_i - \hat{y}_j\right)^2, \tag{6.3}$$

where $\lambda > 0$ is a hyperparameter controlling how strongly smooth across neighbours, and $w_{ij} = w_{ji}$ are symmetric non-negative weights ($w_{ij} = 1$ if i, j share a border or lie within some radius, else 0). Thus, at iteration t, the overall objective (for fitting the next tree f_t) becomes:

$$\mathscr{L}^{(t)} = \mathscr{L}^{(t)}_{\text{data}} + \mathscr{L}^{(t)}_{\text{spatial}} + \mathscr{L}^{(t)}_{\text{reg}}. \tag{6.4}$$

$$\mathscr{L}^{(t)}_{\text{data}} = \sum_{i=1}^{n} \ell\left(y_i, \hat{y}_i^{(t-1)}\right), \tag{6.5a}$$

$$\mathscr{L}^{(t)}_{\text{spatial}} = \frac{\lambda}{2} \sum_{i,j} w_{ij} \left[\left(\hat{y}_i^{(t-1)} + f_t(x_i)\right) - \left(\hat{y}_j^{(t-1)} + f_t(x_j)\right)\right]^2, \tag{6.5b}$$

$$\mathscr{L}^{(t)}_{\text{reg}} = \sum_{k=1}^{t} \Omega(f_k). \tag{6.5c}$$

where $\Omega(f)$ in Eq.(6.5c) is the usual XGBoost tree complexity regularisation for a single leaf f. The data-loss term (decomposes over i) and the spatial-penalty term are defined in Eq.(6.5a) and Eq.(6.5b), respectively.

Then we need to derive the gradient and Hessian. XGBoost's split-finding and leaf-weight formulas rely on two items for each training example i at iteration t. In standard XGBoost loss function, because the data loss $\ell(y_i, \hat{y}_i)$ is separable across i, we have

$$g_i^{(\text{data})} = \left.\frac{\partial \ell(y_i, \hat{y}_i)}{\partial \hat{y}_i}\right|_{\hat{y}_i = \hat{f}_i^{(t-1)}}, \quad h_i^{(\text{data})} = \left.\frac{\partial^2 \ell(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}\right|_{\hat{y}_i = \hat{y}_i^{(t-1)}}.$$

But with the spatial penalty term, $\hat{y}_i$ and $\hat{y}_j$ are joint. So we can compute the gradient of the spatial item to $\hat{y}_i$ (holding $\hat{y}_j$ constant when $j \neq i$). When we differentiate by $\hat{y}_i^{(t-1)}$, the spatial penalty term in Eq.(6.5b) gives

$$\frac{\partial}{\partial \hat{y}_j^{(t-1)}}\left[\frac{\lambda}{2}\sum_j w_{ij}\left(\hat{y}_i^{(t-1)}-\hat{y}_j^{(t-1)}\right)^2\right]=2\lambda\sum_{j=1}^{n} w_{ij}\left(\hat{y}_i^{(t-1)}-\hat{y}_j^{(t-1)}\right).$$

XGBoost's split-finding algorithm does not use off-diagonal Hessian entries (it only expects a diagonal Hessian for efficiency). Putting it all together, for each sample i, the gradient and Hessian contributions are:

Data-loss part

$$g_i^{(\text{data})}=\left.\frac{\partial \ell(y_i,\hat{y}_i)}{\partial \hat{y}_i}\right|_{\hat{y}_i=\hat{y}_i^{(t-1)}},\quad h_i^{(\text{data})}=\left.\frac{\partial^2 \ell(y_i,\hat{y}_i)}{\partial \hat{y}_i^2}\right|_{\hat{y}_i=\hat{y}_i^{(t-1)}}.$$

Therefore, the total gradient and diagonal Hessian to XGBoost at iteration t are:

Gradient

$$g_i^{(t)}=\underbrace{\left.\frac{\partial \ell(y_i,\hat{y}_i)}{\partial \hat{y}_i}\right|_{\hat{y}_i=\hat{y}_i^{(t-1)}}}_{g_i^{(\text{data})}}+\underbrace{2\lambda\sum_{j=1}^{n} w_{ij}\left(\hat{y}_i^{(t-1)}-\hat{y}_j^{(t-1)}\right)}_{g_i^{(\text{spatial penalty})}}.$$

Diagonal Hessian

$$h_i^{(t)}=\underbrace{\left.\frac{\partial^2 \ell(y_i,\hat{y}_i)}{\partial \hat{y}_i^2}\right|_{\hat{y}_i=\hat{y}_i^{(t-1)}}}_{h_i^{(\text{data})}}+\underbrace{2\lambda\sum_{j=1}^{n} w_{ij}}_{h_i^{(\text{spatial penalty})}},$$

where $w_{ij}=w_{ji}, w_{ii}=0$.

6.3.2 K-Nearest neighbours interpolation

The K-Nearest Neighbours (KNN) algorithm is typically used to interpolate values at unobserved locations based on a set of observed data points (Peterson, 2009). Let $\mathscr{D}=\{(x_i,y_i)\}_{i=1}^n$ be the set of n observed samples, where $x_i\in\mathbb{R}^d$ denotes the spatial location and $y_i\in\mathbb{R}$ denotes the measurement. For an unobserved target location $x^\star$, KNN interpolation proceeds in two main steps:

Neighbours selection

Compute the Euclidean distance from $x^\star$ to each observed point:

$$d(x^\star,x_i) \;=\; \|x^\star-x_i\|_p, \tag{6.6}$$

where $\|\cdot\|_p$ is the ℓ_p norm (commonly $p=2$ for Euclidean distance). Sort the distances and select the K points with the smallest $d(x^\star,x_i)$. Let $i_{(1)},i_{(2)},\ldots,i_{(n)}$ be a permutation of $\{1,\ldots,n\}$

such that

$$d_{i_{(1)}} \leq d_{i_{(2)}} \leq \cdots \leq d_{i_{(n)}}.$$

Then the K nearest neighbours are

$$\mathscr{N}_K(x^\star) = \{\, i_{(1)}, i_{(2)}, \ldots, i_{(K)} \}.$$

With these neighbours, the interpolated value is

$$\hat{y}^\star = \frac{\sum_{i \in \mathscr{N}_K(x^\star)} w_i \, y_i}{\sum_{i \in \mathscr{N}_K(x^\star)} w_i} \tag{6.7}$$

$$w_i \;=\; \frac{1}{(d(x^\star, x_i) + \varepsilon)^\alpha}, \tag{6.8}$$

where $\alpha > 0$ controls how rapidly influence decays with distance (often $\alpha = 2$), and $\varepsilon > 0$ is a small regulariser to avoid division by zero when $x^\star$ coincides with a sampled location.

6.3.3 Conformal prediction

Conformal prediction wraps any model to give distribution-free prediction sets with a chosen coverage level, using a held-out calibration set. This assumes the data are exchangeable. Given data $Z_i = (x_i, y_i)$, $i = 1, \ldots, n$, and a new covariate vector x_{n+1}, the aim is to construct a distribution-free $(1-\alpha)$ prediction set for y_{n+1}, where α is the miscoverage rate. To be specific, split the data into a training set $\mathscr{D}_{\text{train}}$ and a disjoint calibration set $\mathscr{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^{m}$, where m denotes the number of calibration scores. Fit a predictor $\hat{f}$ on $\mathscr{D}_{\text{train}}$ and compute residual scores on calibration:

$$S_i = \left| y_i - \hat{f}(x_i) \right|, \qquad (x_i, y_i) \in \mathscr{D}_{\text{cal}},\ i = 1, \ldots, m.$$

From the finite-sample quantile

$$\widehat{q}_{1-\alpha} = \inf\left\{ t \;:\; \frac{\#\{\, i : S_i \leq t \,\} + 1}{m+1} \geq 1 - \alpha \right\}.$$

The $(1-\alpha)$ conformal prediction set for x_{n+1} is

$$C_\alpha(x_{n+1}) = \left\{ y :\ |y - \hat{f}(x_{n+1})| \leq \hat{q}_{1-\alpha} \right\} = \left[\hat{f}(x_{n+1}) - \hat{q}_{1-\alpha},\ \hat{f}(x_{n+1}) + \hat{q}_{1-\alpha} \right].$$

If the calibration examples and the test point are exchangeable, then

$$\mathbb{P}\big\{ Y_{n+1} \in C_\alpha(X_{n+1}) \big\} \;\geq\; 1 - \alpha.$$

6.4 XGBoost experiment design

To systematically evaluate XGBoost for spatio-temporal prediction and ensure reproducible results, the experimental design includes four parts: data preparation (to control data quality), feature engineering (to encode the general predictor and spatio-temporal structure that the XGBoost learner does not model directly), prediction setup (to set the steps for how to make the prediction properly), and validation strategy (to obtain unbiased model performance and coverage estimates under spatial and temporal dependence).

6.4.1 Data preparation for XGBoost

Like other gradient-boosting model frameworks, XGBoost assumes a fully complete feature matrix (no missing elements in the matrix) to build decision tree ensembles efficiently without introducing spurious splits. Although XGBoost can handle missing values in its tree-building process (Chen, 2016), this approach is particularly effective when the missingness itself carries meaningful information. However, in most environmental sensor networks, data gaps typically come from device failure or maintenance rather than a hidden pattern. Chen (2016) explains how XGBoost handles missing values during the construction of decision trees. While XGBoost can assign a default direction for missing values at each split, the most reliable approach is to handle these missing values before constructing the tree. This is recommended to avoid treating missing values as a separate branch, which can negatively impact performance if the missingness is not naturally informative. Therefore, it is often recommended to impute or interpolate missing values before modelling (APXML).

In real-world scenarios, spatio-temporal environmental datasets, such as soil moisture readings collected from multiple sensors, often contain missing values (e.g., Figure 6.3). This can happen due to device failures, maintenance issues, or errors on specific days. When a single sensor fails to provide a reading at a given time point, it creates a gap in the time series at that location. This missing data also disrupts the uniform (location $\times$ time) panel table needed for calculating lag terms and neighbour-based features.

To prepare the data for XGBoost, we need to create a comprehensive panel table that explicitly lists every (location $\times$ time) pair across the entire period. This table will indicate any missing values, followed by temporal interpolation and edge-filling methods for estimating those missing data points at the two boundaries. It's essential to ensure that each sensor site provides consistent inputs every day during the modelling window.

We categorise missing values into two cases based on the duration of the missing data. If a sensor has gaps of three consecutive days or fewer, we estimate the missing measuring values using cubic interpolation (Lam, 1983). If it has more than three consecutive days with missing values,

then those days are excluded from the modelling and are skipped. Cubic interpolation is chosen because it is suitable for environmental data and has several benefits. Unlike linear interpolation, which focuses on sharp changes at every observation, cubic spline interpolation ensures that both the first and second derivatives are continuous across different observations. Since the environmental data, such as soil moisture and other environmental variables, often change gradually over time, a twice differentiable approach can more accurately reflect the underlying physical processes. In addition, real environmental processes rarely change at a constant rate. Therefore, a piecewise cubic polynomial can be flexible enough to capture the local increases or decreases in trends, resulting in a more realistic curve rather than straight segments.

Furthermore, for gaps with small intervals, cubic splines have flexibility without overfitting. This means they can smooth out noise while keeping important dynamics. Lastly, solving the cubic spline is computationally efficient, with a complexity of $\mathcal{O}(n)$ where n is the size of the input.

For each interval $[t_i, t_{i+1}]$, we fit a natural cubic spline defined as

$$S_i(t) = a_i(t-t_i)^3 + b_i(t-t_i)^2 + c_i(t-t_i) + d_i,$$

ensuring continuity in the function and its first two derivatives at each knot, along with natural boundary conditions where $S''(t_0) = S''(t_n) = 0$ (Keys, 1981). Any sensor that shows a single continuous gap longer than three days will be removed from further modelling, as the longer gap cannot be reliably interpolated and may introduce bias. The training set and test set are always split by date before the imputation. Then, all imputation parameters are learned on the training set and passed to the test set to avoid data leakage. In LOOCV, we repeat the same procedure within each fold to ensure that no future data leaks into the model training.

In summary, while XGBoost can handle missing values by assigning a default direction during splits, spatio-temporal sensor readings frequently have gaps. Therefore, creating a complete data panel and filling in missing values is an essential step for developing a reliable XGBoost spatio-temporal model. With all features available for each sample, XGBoost can focus on learning real relationships, such as how today's soil moisture depends on yesterday's readings and the spatial correlations, without worrying about arbitrary missing values in the dataset.

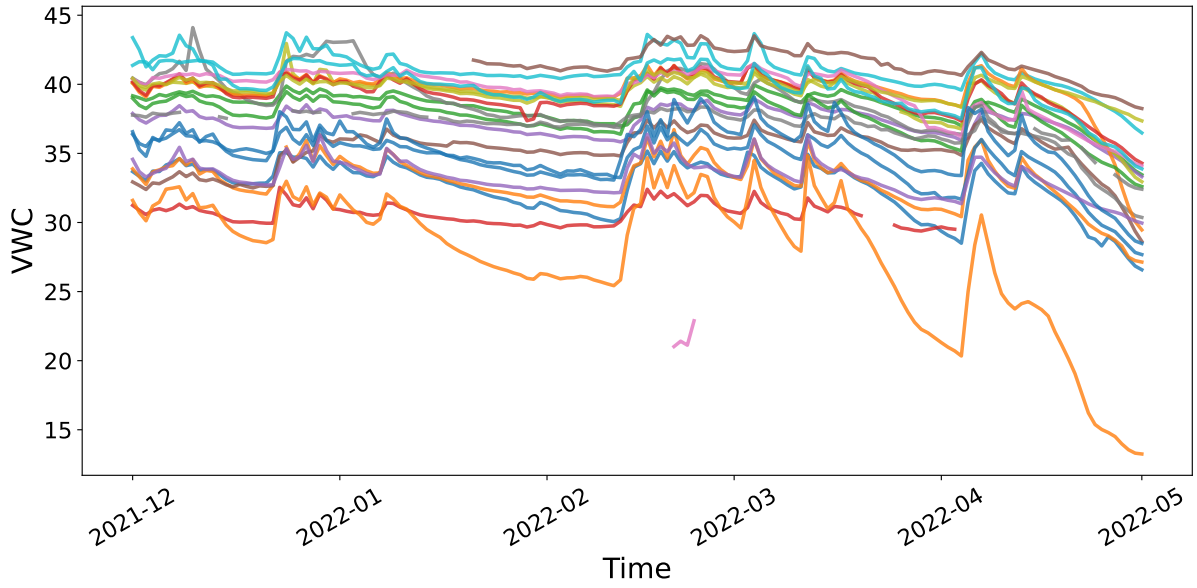


Figure 6.1: Time series of daily volumetric water content (VWC) for all soil moisture sensors from 2022-01-01 to 2022-05-28.

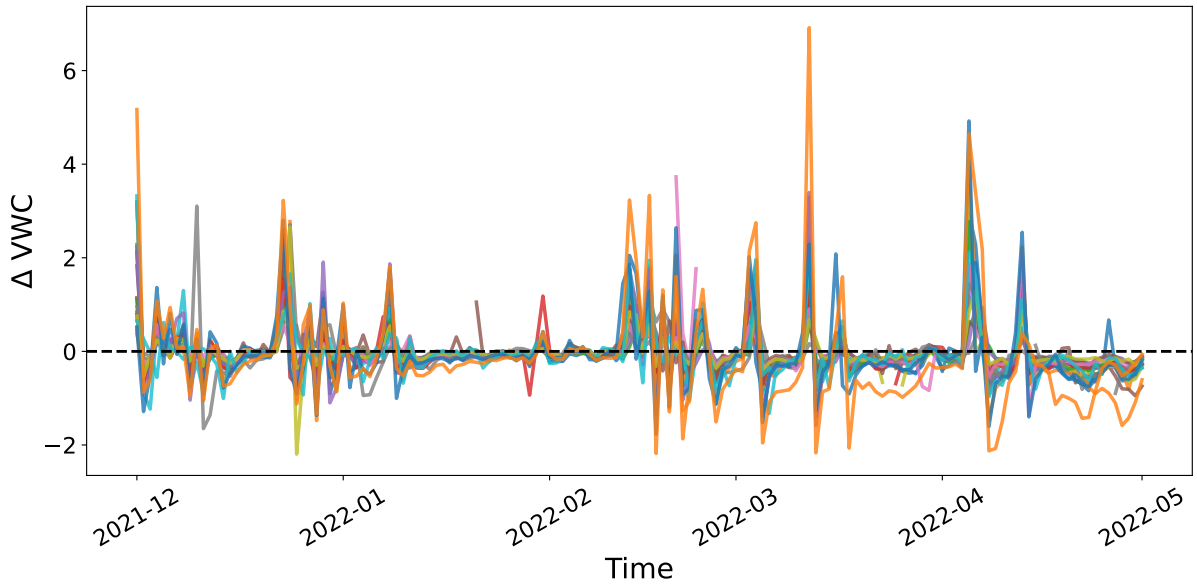


Figure 6.2: Time series of daily change in VWC (Δ VWC) for all sensors from 2022-01-01 to 2022-05-28

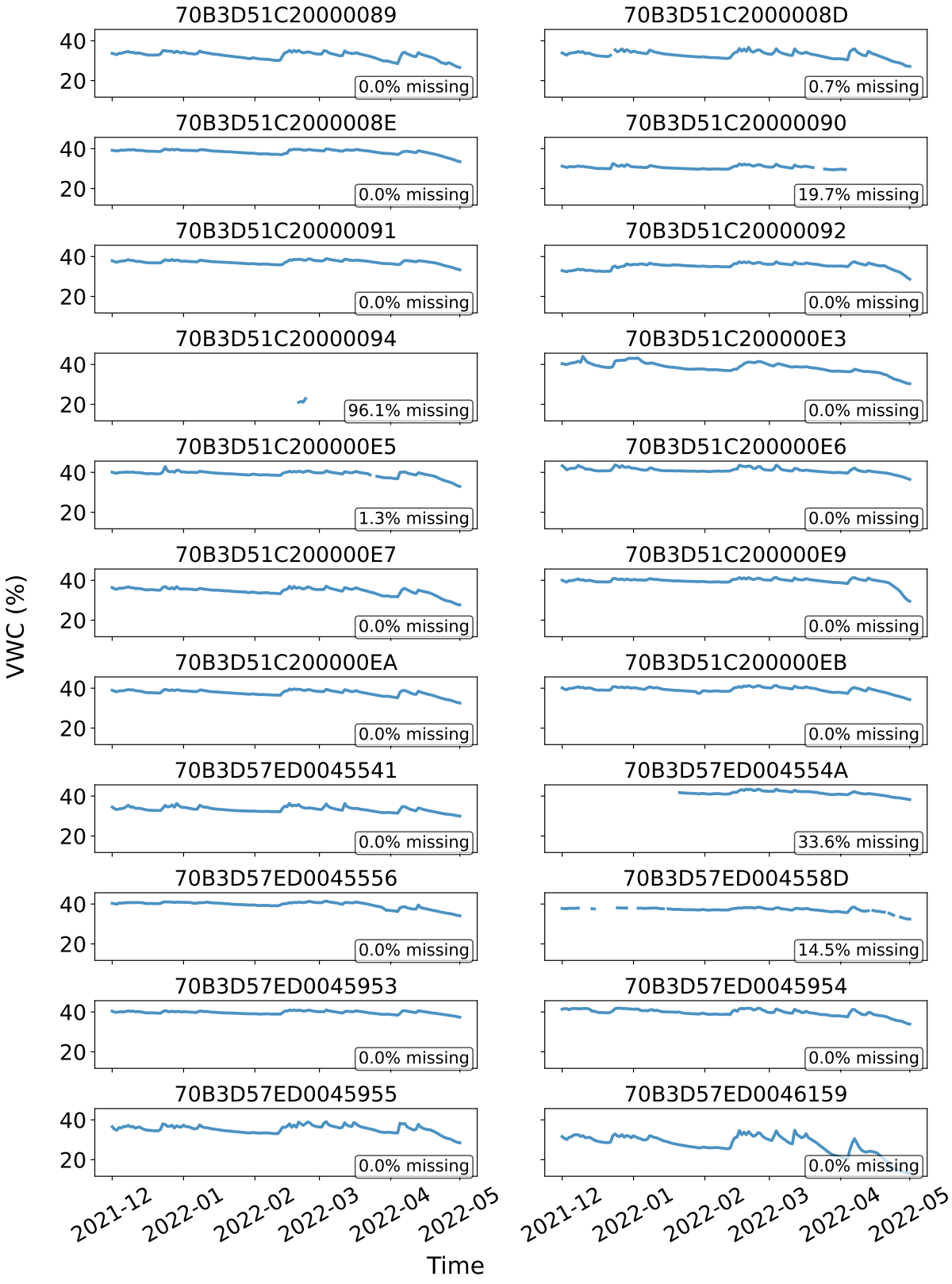


Figure 6.3: VWC time series for all sensors with a missing percentage from 2022-01-01 to 2022-05-28

Figure 6.3 shows the VWC time series for all sensors with the missing percentage. After data

preparation, Figures 6.1 and 6.2 present the time series of volumetric water content (VWC) and its daily change (Δ VWC) for all sensors, spanning from January 1 to May 28, 2022.

6.4.2 Predictor construction (spatio-temporal features)

We transform the variables in the soil moisture dataset into predictors and fuse multiple sources, including in-situ sensors (VWC), satellite soil moisture product (SWI), meteorological variables, and elevation, to maximise XGBoosts model power for fine-resolution mapping. We then transform features that the model can use: encode categorical variables, build local VWC summaries (e.g., KNN), and create lagged rainfall covariates. These spatial and non-spatial features help the trees learn cross-variable relationships and yield a high-resolution soil-moisture model that leverages the strengths of each source.

The XGBoost model inputs combine multiple sources of data:

- **Satellite data:** Soil Water Index (SWI) at time t, available on a coarse grid.
- **Satellite patch:** the 3×3 neighbourhood around the sensors grid cell, flatten into an array to make fully use of the rich spatial satellite data.
- **Meteorological data:** Daily rainfall (both current and lag-1), and soil temperature from nearby weather stations.
- **Topographical data:** Elevation at each prediction location.
- **In-Situ observations:** VWC from ground sensors, used for training and evaluation.
- **Spatial features:** Easting and northing coordinates to capture spatial trends.
- **Local VWC:** To incorporate neighbourhood information into the XGBoost model, we compute a local VWC feature by averaging the readings of the four nearest sensors surrounding the target sensors.

Table 6.1: summary of input spatial and aspatial features used for soil moisture prediction

Category	Features
In-situ sensor observations	Volumetric Water Content (VWC), Soil Temperature
Satellite images	Soil Water Index (SWI)
Meteorological covariates	Rainfall (current day), Rainfall (lag-1 day)
Topographical variables	Elevation
Geographic coordinates	Easting, Northing, Local VWC

All spatial and aspatial features are interpolated onto a common prediction grid of size 100×100 using KNN (with $k = 4$) for fine-resolution map prediction. The XGBoost target variable is

the daily change in VWC, namely, the difference in VWC between two consecutive days. The target is $\Delta\text{VWC}_t = \text{VWC}_{t+1} - \text{VWC}_t$. We model the daily change ΔVWC_t because it captures short-term dynamics and is less affected by level bias. Since day-to-day levels often vary little, which makes direct level modelling harder, so we predict the daily change and recover the next day via $\text{VWC}_{t+1} = \text{VWC}_t + \Delta\text{VWC}_t$.

6.4.3 Prediction setup

The aim is to predict next-day VWC. The model is trained on features on day t and the final prediction is obtained by applying the reverse transform $\text{VWC}_{t+1} = \text{VWC}_t + \Delta\text{VWC}_t$. The training set is constructed using a rolling multi-day window to increase the sample size and capture temporal variability. All features are interpolated to fine-resolution features using KNN, and these KNN-interpolated features are then passed into the trained XGBoost ensemble. Each decision tree in the ensemble applies its learned splits to generate a soil moisture estimate at every fine-resolution point. By combining the spatial continuity offered by KNN with XGBoost's ability to model complex, non-parametric interactions, the fine-resolution map reflects the local variability and relationships captured during XGBoost training.

6.4.4 Hyperparameter tuning and validation strategy

To assess the predictive performance of the XGBoost data fusion model, we use two complementary validation strategies and select the spatial-smoothing hyperparameter λ by cross-validation. First, for temporal robustness, we apply rolling time series cross-validation: the model is trained on past windows and validated on subsequent periods to ensure it captures evolving dynamics across time. Second, for spatial generalisation, we use leave-one-sensor-out cross-validation. In each iteration, one in-situ sensor is held out during training and used only for testing, which evaluates the models ability to predict at unseen locations.

6.4.4.1 Rolling time series cross-validation

We use a rolling split: 30 days for training, 14 days for calibration, and 1 day for testing across the study period.

$$t_k \in \{\text{2022-01-31}, \ldots, \text{2022-05-28}\}.$$

For each test day t_k, we train the model on the previous 30 days and then predict on day t_k. We slide the window forward one day at a time and repeat for every single day in the selected period.

$$\underbrace{\{t_k - 30, \ldots, t_k - 15\}}_{\text{training (30 days)}} \Big| \underbrace{\{t_k - 14, \ldots, t_k - 1\}}_{\text{calibration (14 days)}} \longrightarrow \underbrace{\{t_k\}}_{\text{test}}$$

We compute the one-day-ahead prediction $\hat{y}_{t_k}$ and save the fold RMSE:

$$\mathrm{RMSE}_k(\lambda) = \sqrt{\left(y_{t_k} - \hat{y}_{t_k}\right)^2}.$$

Aggregating over N folds gives

$$\mathrm{CV\_RMSE}(\lambda) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(y_{t_k} - \hat{y}_{t_k}^{(\lambda)}\right)^2}.$$

This rolling CV approach considers the temporal pattern and avoids data leakage (Bergmeir and Benítez, 2012).

Global λ selection

λ in Eq. (6.3) controls the strength of the spatial-smoothing penalty. It balances data fitting and smoothness. For the final model, we selected a single global regularisation weight, denoted as λ^*, by minimising the average one-day-ahead RMSE across all rolling windows instead of tuning separately for each day. Specifically, for each candidate λ from a grid defined as

$$\Lambda = \{10^{-4}, 10^{-3}, \ldots, 10^{2}\},$$

We compute a cross-validation score using the 30-3-1 split

$$\underbrace{\{D-33,\ldots,D-4\}}_{\text{train (30 d)}} \;\Big|\; \underbrace{\{D-3,D-2,D-1\}}_{\text{validation (3 d)}} \longrightarrow \underbrace{\{D\}}_{\text{test (1 d)}},$$

and define

$$\mathrm{CV\_RMSE}(\lambda) = \sqrt{\frac{1}{\sum_k |V_k|}\sum_k \sum_{t\in V_k}\left(y_t - \hat{y}_t^{(\lambda)}\right)^2}, \qquad V_k = \{D_k-3, D_k-2, D_k-1\}.$$

We then choose

$$\lambda^* = \arg\min_{\lambda\in\Lambda} \mathrm{CV\_RMSE}(\lambda).$$

We first select a single global λ^* on an earlier calibration period (2021-12-01 to 2022-1-31) and then fix λ^* for all following one-day-ahead predictions. With λ^* fixed globally, we retrain each 30-day model using this value and produce the one-day-ahead predictions for day D for the whole study period. This strategy reduces overfitting in hyperparameter selection and improves reproducibility (Cawley and Talbot, 2010), while ensuring that no test-day information is used for tuning. By fixing λ^* globally, we maintain consistent smoothing and reduce the risks of instability and overfitting associated with tuning on a per-day basis, while still capturing the spatial penalty that enhances average prediction accuracy compared to the default loss.

Per window λ tuning

An alternative way to choose λ is to pick

$$\lambda_k^* = \arg\min_{\lambda \in \Lambda} \mathrm{RMSE}_k^{\mathrm{val}}(\lambda), \quad \mathrm{RMSE}_k^{\mathrm{val}}(\lambda) = \sqrt{\frac{1}{|V_k|} \sum_{t \in V_k} \left(y_t - \hat{y}_t^{(\lambda)}\right)^2},$$

independently for each fold k. However, this validation approach

- increases computational cost,
- may overfit to the noise in the short time window,
- slightly complicates the comparison across multiple days.

6.4.4.2 Leave one sensor out cross-validation

To evaluate spatial generalisation and prevent overfitting to a specific sensor, we enhance the temporal cross-validation with a leave-one-sensor-out (LOSO) approach:

1. Let $\mathscr{S} = \{1, \ldots, p\}$ index the set of p sensors.
2. For each sensor $s \in \mathscr{S}$:
   (a) Remove all observations from sensor s from the training data.
   (b) Train the XGBoost model on the remaining $p-1$ sensors, using the same 30-day rolling windows, filling gaps ≤ 3 days by cubic splines and dropping any window with a gap > 3 days.
   (c) Produce one-day-ahead prediction for sensor s over its available dates in the evaluation period (2022-01-01 to 2022-05-28).
   (d) Compute the sensor-specific RMSE:

$$\mathrm{RMSE}_s = \sqrt{\frac{1}{N_s} \sum_{i=1}^{N_s} \left(y_{s,i} - \hat{y}_{s,i}\right)^2},$$

   where N_s is the number of heldout days for sensor s.
3. Aggregate performance across sensors:

$$\mathrm{LOSO\_RMSE} = \frac{1}{p} \sum_{s=1}^{p} \mathrm{RMSE}_s,$$

This LOSO-CV evaluates how well the model can predict at an entirely unknown sensor location, to test the spatial bias introduced by the Laplacian penalty. It has been used in environmental sensor network studies to quantify spatial transferability (Cressie, 1993).

These two validation strategies comprehensively assess both the temporal and spatial stability of high-resolution soil moisture predictions.

6.5 XGBoost point data and grid data fusion

This experiment estimates soil moisture over the catchment by integrating sparse in-situ sensors, satellite gridded data, and environmental covariates (e.g., rainfall, temperature). The main prediction task here can be formulated as a supervised regression problem, where the target is the volumetric water content (VWC) at each spatial location. We use XGBoost and encode spatial structure by adding a spatial-smoothness penalty to the loss function so neighbouring locations have smoother predictions. The uncertainty is quantified via spatio-temporal, locally scaled conformal prediction. Both same-day (t) and one-day-ahead ($t+1$) predictions are considered in this experiment.

6.5.1 Implementation of the XGBoost's custom loss function

XGBoost's standard objective in the loss function does not account for spatial autocorrelation among sensors. To borrow strength from nearby sensor measurements, we modify the loss function with a spatial penalty (See Section 6.3.1 for details). The steps are:

1. **Define neighbour relations.** For each sensor i, find its K nearest neighbours by geographic distance. We form a binary adjacency matrix A, which is widely used in many spatial models (Cliff and Ord, 1981).

$$A_{ij} = \begin{cases} 1, & \text{if sensor } j \in \mathcal{N}_i, \\ 0, & \text{otherwise,} \end{cases}$$

   where $\mathcal{N}_i$ is the set of the K closest sensors to i.

2. Weight by distance. However, the sensors lie on an irregular, sparse network rather than a grid, encoding neighbour relations only based on the adjacency matrix, which ignores the information on the distance between sensors. Therefore, we not only consider the adjacency but also include the distance of different neighbours. To make the actual Euclidean neighbour distances d_{ij} influence, we encode them into W_{ij}. For example:

$$W_{ij} = \begin{cases} \dfrac{1}{d_{ij}+\varepsilon}, & j \in \mathcal{N}_K(i), \\ 0, & \text{otherwise}, \end{cases}$$

$$w_{ij} = A_{ij}\frac{1}{d_{ij}+\eta}, \qquad d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|,$$

with $\eta > 0$ small to avoid division by zero.

3. **Normalise to a matrix.** Let

$$D_i = \sum_{j=1}^{N} w_{ij}, \quad \tilde{w}_{ij} = \frac{w_{ij}}{D_i}.$$

In matrix form, if $W = [w_{ij}]$ then $\tilde{W} = D^{-1}W$ with $D = \mathrm{diag}(D_i)$.

$$\mathscr{L}(f) = \sum_i (f_i - y_i)^2 + \lambda \sum_{i,j} w_{ij}(f_i - f_j)^2 + \sum_{k=1}^{K} \Omega(f_k).$$

4. **Add a spatial penalty to the loss.** Denote by $\mathbf{p} \in \mathbb{R}^N$ the vector of raw tree ensemble predictions (one per sensor). We add

$$\mathscr{L}_{\text{spatial}} = \frac{\lambda}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \tilde{w}_{ij}(p_i - p_j)^2 = \frac{\lambda}{2}\mathbf{p}^\top(\mathbf{I} - \tilde{W})\,\mathbf{p},$$

where $\lambda \geq 0$ controls the strength of spatial smoothing.

Nearby sensors share rainfall and soil properties, so their residuals should be similar. Averaging across neighbours filters out sensor noise. If one sensor fails, its prediction is drawn toward its neighbours rather than relying solely on its past.
To ensure the robust estimation of the penalty term λ in the customised loss function, we compare two alternative tuning schemes (Details are provided in Section 6.4.4):

1. **Global** λ: choose a single λ once, using all rolling windows.
2. **Per window** λ: retune λ separately for each 30 day training window.

6.5.2 Validation strategy 1: Cross-validation on multiple time points

Soil moisture exhibits complex temporal dynamics, and as its variability shifts over time, the model performance accuracy also rises and falls across different test time points. To thoroughly evaluate how the customised loss function performs under these real-world conditions, which are characterised by measurement noise, sudden hydrological changes, and seasonal patterns, a

data-based cross-validation scheme is employed. Specifically, a whole month of data is selected as the test set. By comparing performance metrics (e.g., RMSE and R^2) across these test time points, we can quantify the loss function's ability to learn varying variance and quantify the impact of sudden changes.

Assess the temporal (marginal) validity and stability of our spatio-temporal conformal procedure across a sequence of test days, rather than on a single date. This shows both the long-term coverage level and its daily variability in non-stationary weather conditions. All evaluation metrics are computed on the test set only. This strategy targets marginal validity over time; per-site conditional validity is assessed separately in the per-sensor scheme.

6.5.2.1 Naive model

The naive baseline assumes tomorrow's VWC equals today's VWC, providing a simple error metric that any prediction model must beat. We use RMSE to measure the average size of one-step-ahead errors. The baseline's RMSE thus sets a threshold: a more advanced model must be below it to add value. By comparing the RMSE for both the XGBoost model and the naïve baseline on the same test data, we establish a clear performance floor: only a model with a lower RMSE can be said to capture real soil moisture patterns. To be specific, we can plot the error distribution of each method and examine its performance by sensor (EUI) or environmental condition. This reveals where the data fusion model truly outperforms persistence, guiding further feature engineering and hyperparameter tuning. The RMSE for the naïve one-step error are defined as follows:

Naïve model for the temporal cross-validation

In the naïve benchmark, tomorrows soil moisture is assumed to be equal to todays:

$$\hat{y}_{t+1}^{\text{naive}} = y_t.$$

The one-step-ahead forecast error is

$$e_{t+1} = y_{t+1} - \hat{y}_{t+1}^{\text{naive}} = y_{t+1} - y_t,$$

The prediction accuracy is measured by the RMSE:

$$\text{RMSE}_{\text{naive}} = \sqrt{\frac{1}{N-1}\sum_{t=1}^{N-1}\left(y_{t+1} - y_t\right)^2}. \tag{6.9}$$

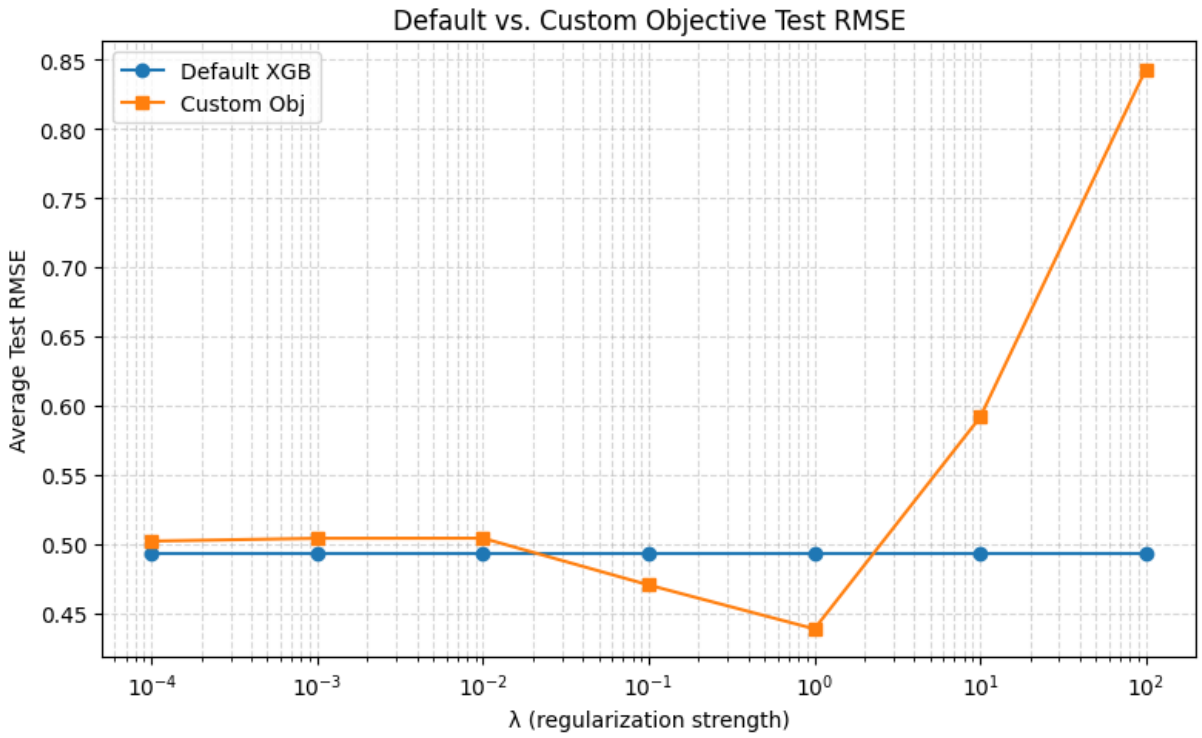


Figure 6.4: Selected global regularisation parameter λ used for all sensors from 2022-01-01 to 2022-05-28 under temporal cross-validation. The same value of λ is applied across all training windows and sensor locations, providing a baseline with a fixed degree of spatial smoothing.

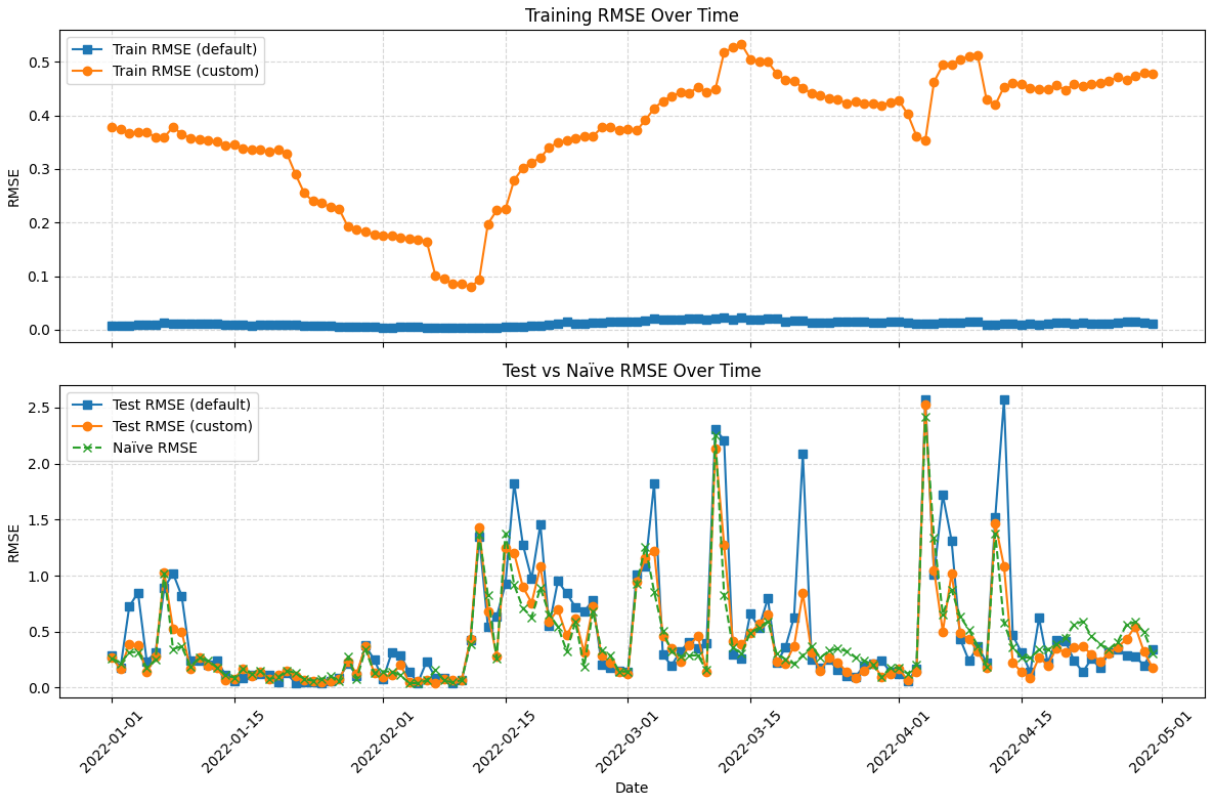


Figure 6.5: Temporal cross-validation using a single global regularisation parameter λ shared across all sensors, for the period from 2022-01-01 to 2022-05-28. This setup assumes a constant degree of spatial smoothing over time and across the network, providing a baseline for comparison with schemes that allow λ to vary by window or sensor.

Figure 6.5 presents the RMSE of XGBoost models trained with both the default and a custom loss function. The custom loss incorporates a global λ chosen by minimising the average one-day-ahead RMSE across all rolling windows (see Figure 6.4 as an example). At the very small λ values, the spatial penalty is switched off, so the two XGBoost models are identical, and the RMSE of XGBoost with the default loss is similar to the RMSE of XGBoost with the customised loss function. It is noted that the RMSEs are not completely the same, possibly because of randomness or the way the custom loss is computed. As λ increases to the moderate range, the customised loss function begins to penalise rapid spatial change of the target variable, which somehow suppresses over-fitting to the noise and drives the customised loss RMSE below the default loss of the RMSE. However, if the λ increases to a very high value, the model pushes the smoothing to the predictions, which introduces bias and causes the RMSE to increase again. The RMSE shows a very classic U-shape curve: a steep plateau at low λ, a lowest bottom point where the regularisation is absolutely right, and a sharp rise again at large λ where the under-fitting dominates the training process. At the same time, the default loss function curve remains flat across all the λ, since it does not consider the spatial penalty. In the upper panel (training set) of Figure 6.5, the RMSE for the model with the custom loss fluctuates widely, showing that a single global λ cannot optimise prediction performance for every single day. It trades some pointwise

accuracy for regularisation, to reduce overfitting and improve generalisation. In the lower panel (test set) of Figure 6.5, however, the custom loss model often outperforms the default loss model, achieving lower RMSE values and confirming that XGBoost gains predictive benefit from the customised objective from the temporal perspective.

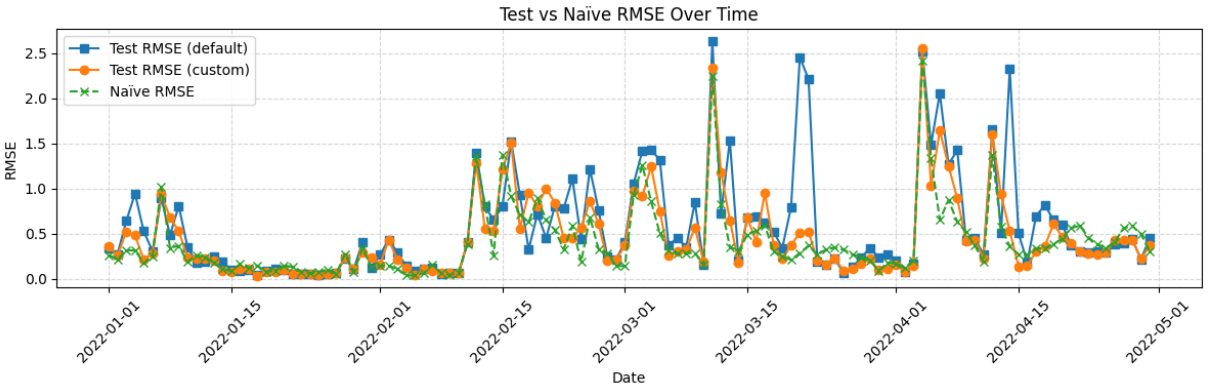


Figure 6.6: Temporal cross-validation using different λ for each window for all sensors from 2022-01-01 to 2022-05-28

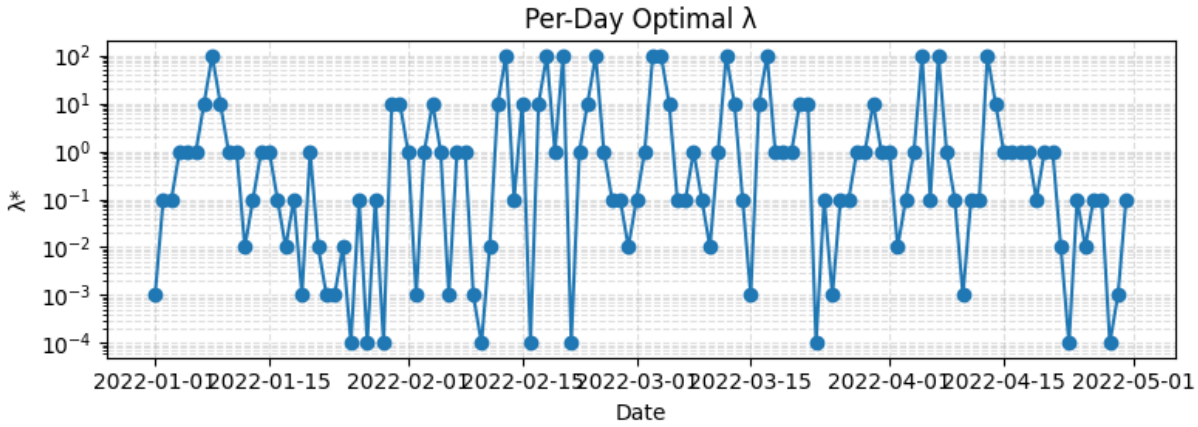


Figure 6.7: λ for temporal cross-validation using different λ for each window for all sensors from 2022-01-01 to 2022-05-28

Figure 6.6 shows the RMSE for the custom loss and default loss, and the naive RMSE with the best λ for each time window. Figure 6.7 shows the best λ for every time window for each test day. The λ varies a lot from day to day, which shows the sudden change and non-stationarity of the soil moisture data. The RMSE of the customised loss function XGBoost is constantly smaller than the RMSE of the default loss function, which means the custom loss (extra regularisation) helps to smooth out the noise.

We evaluate the XGBoost model in an out-of-sample setting using a rolling 30-day training window to generate a one-day-ahead prediction for each day in the period January 2022 to May 2022. Specifically, for each test day t, we train on the immediately preceding 30 days

$\{t-30,\ldots,t-1\}$. Because the number of active sensors can be missing at different time points (in Section 2.1), we use the following data quality filter on each time window:

1. Drop the sensor entirely within that window if any gap of missing observations exceeds three days.

2. Otherwise, fill all gaps of up to three consecutive days using cubic-spline interpolation.

After applying the data quality filter, we obtain valid one-day-ahead predictions for each $t \in \{2022\text{-}01\text{-}01,\ldots,2022\text{-}05\text{-}28\}$. We then compute the test-set RMSE as

$$\mathrm{RMSE}_{\mathrm{naive}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - y_{i-1}\right)^2}$$

as per equations (6.9). Figure 6.6 shows that for most test days, the model's RMSE lies below the naïve RMSE, which demonstrates consistent outperformance of the baseline predictor.

6.5.3 Validation strategy 2: Leave one sensor out cross-validation

To further assess the loss function on the spatial generalisation, we conduct a leave-one-sensor-out cross-validation on a single day that captures the spatial variation. For each sensor, we remove its value and train the model on the remaining sensors. It is noted that although spatial cross-validation is usually used to account for the fact that close data points show similarity to each other than distant points to avoid overfitting of the spatial autocorrelation for the spatial datasets (Roberts et al., 2017), it is not always necessary to do so consider that we only got 22 (maximum, but most of the time at least one of them is missing) sensors. They are so sparse that even the spatial autocorrelation is hard to spot from them. In other words, if we spread out so widely that there isn't a clear spatial structure to exploit, if we still try to form any spatial blocks, it will end up with just a few folds of uneven size, and each fold will throw away most of the data, which will make it even harder to tune the λ and get stable and reliable estimates. By using the leave-one-sensor-out method instead, we still test on an unseen location but retain as much data as possible in each training set, providing more trustworthy measures of how the penalty term behaves at each site. The main goal here is to figure out how each sensor reacts to the penalty term λ and to test model predictions at entirely unseen locations. Within each fold, we split those training sensors into a sub-training set and a small validation set to tune the penalty term λ hyperparameter of our custom loss function, selecting the value that minimises validation RMSE. With λ fixed, we retrain all the remaining sensors and predict soil moisture at the held-out location, then calculate the root mean squared error (RMSE) against the true values. Repeating this for every sensor yields an optimal λ per sensor location, revealing how noisier or more dynamic locations demand different regularisation strengths and measures how accurately the loss function supports predictions at an entirely unseen point. This method not only uncovers the spatial sensitivity of λ but also demonstrates the loss function's ability to maintain robust

performance across the full sensor network.

We tune the spatial smoothness penalty parameter by leave-one-sensor-out cross-validation (LOSO-CV), selecting a single global value $\lambda^\star$ and using it for all the other sensors. It is noted that the neighbour matrix $\mathbf{W}$ is rebuilt inside each fold using only training sensors to avoid using the held-out sensors information. Rows of W are normalised to sum to one, and the diagonal is zeroed, ensuring the penalty scales comparably across sensors. Because the penalty trades bias for variance, training RMSE typically increases with λ, while test RMSE follows a characteristic U-shape; the chosen $\lambda^\star$ lies near the curve.

Naïve model for the spatial cross-validation

In the spatial naïve benchmark, the soil moisture at the heldout sensor is assumed to equal the value interpolated from neighbouring sensors at the same time:

$$\hat{y}_s^{\text{naive}} = y_{s_{\text{interp}}}.$$

The prediction error is

$$e_s = y_s - \hat{y}_s^{\text{naive}} = y_s - y_{s_{\text{interp}}},$$

And its accuracy is measured by the RMSE:

$$\text{RMSE}_{\text{naive}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - y_{i_{\text{interp}}}\right)^2}.$$

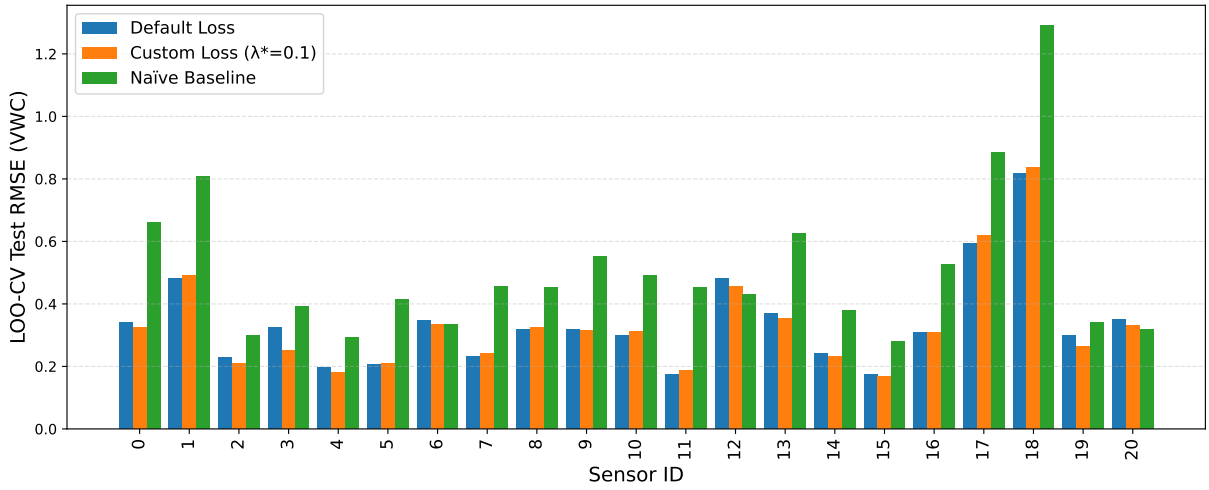


Figure 6.8: Leave-one-sensor-out cross-validation RMSE at each sensor location, comparing three settings: (1) **Default XGBoost**, which uses the built-in loss function and default hyperparameters; (2) **Global λ**, which applies a custom loss with a single, fixed regularization parameter λ chosen to minimize the average error across all days and sensors; and (3) **Naive baseline**, a simple model that predicts the training mean for each day. The default-loss XGBoost achieves the lowest RMSE at almost every site, whereas the global-λ model performs on par with the naive baseline, indicating that one fixed λ cannot accommodate the spatial sparsity of our network.

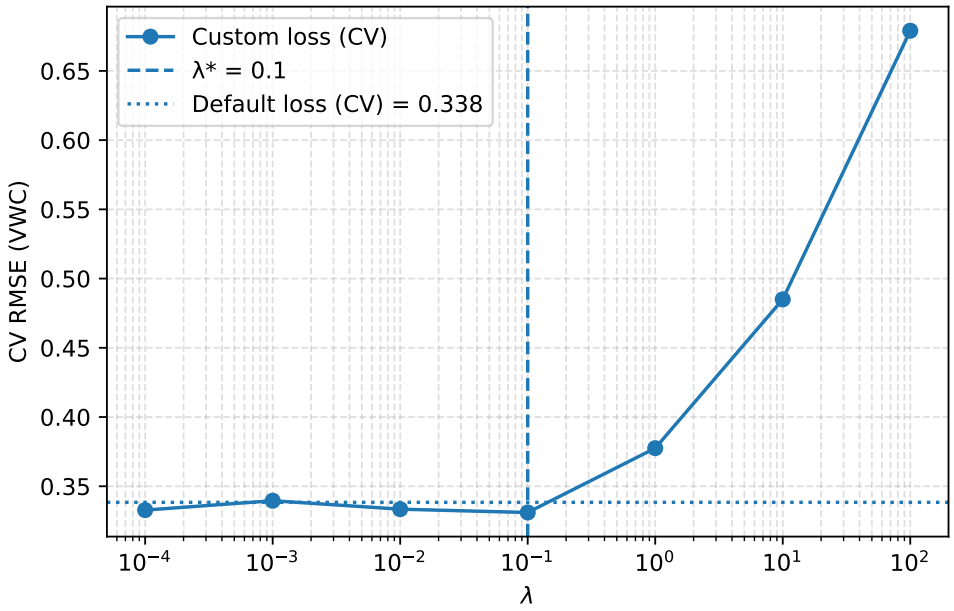


Figure 6.9: Cross-validated RMSE (VWC) of the custom spatially regularised loss as a function of the Laplacian weight λ. The vertical dashed line marks the selected λ^*, and the horizontal dotted line shows the baseline RMSE from the default loss. Small positive λ improves accuracy, while large λ over-smooths and degrades performance.

Based on the RMSE on held-out data shown in Figure 6.8, the default loss function XGBoost model achieves the lowest error at nearly every sensor location, indicating better predictive performance. In contrast, the global custom loss model shows little improvement in RMSE

over the naïve baseline, which shows limited benefits from the global λ setting. This suggests that our sensor network is too spatially sparse: removing any one sensor changes the local data distribution dramatically, and enforcing a single global λ fails to explain these localised changes. When the same λ is applied for the whole cross-validation, the model is either under-regularised at some sites or over-regularised at others, degrading every single sensor prediction accuracy.

To allow location-specific regularisation, we remove all data from the held-out sensor i and tune λ using only the remaining sensors via validation:

$$\lambda_i^* = \arg\min_{\lambda \in \Lambda} \frac{1}{K_i} \sum_{k=1}^{K_i} \mathrm{RMSE}_{i,k}^{\mathrm{val}}(\lambda), \qquad \mathrm{RMSE}_{i,k}^{\mathrm{val}}(\lambda) = \sqrt{\frac{1}{|V_{i,k}|} \sum_{t \in V_{i,k}} \left(y_t - \hat{y}_t^{(\lambda)}\right)^2},$$

where each $V_{i,k}$ is a validation block drawn from sensors $\neq i$. We then refit on the non-i data with λ_i^* and evaluate once on sensor i.

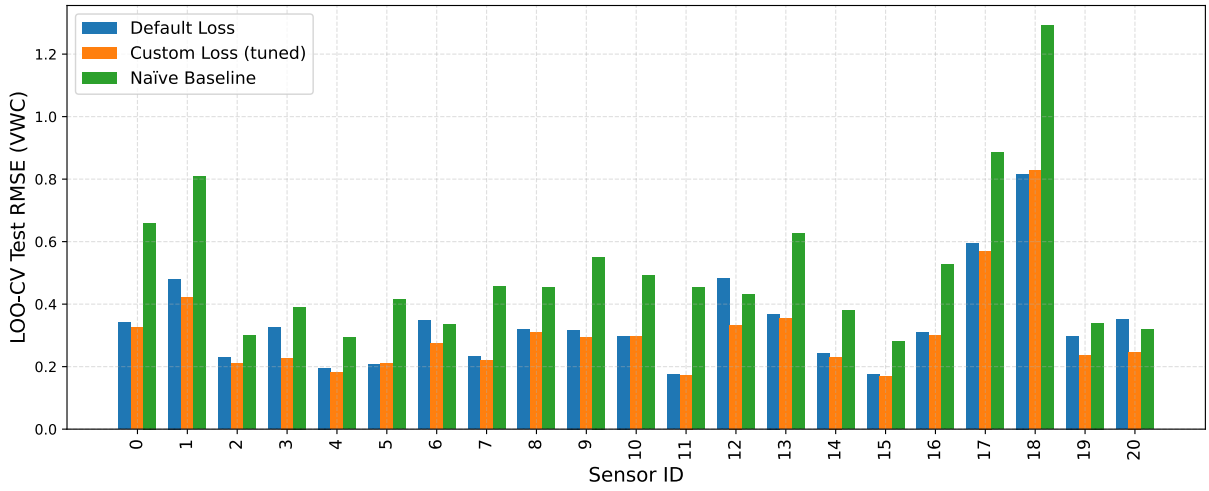


Figure 6.10: Leave-one-sensor-out cross-validation RMSE at each sensor location, comparing three settings: (1) **Default XGBoost**, which uses the built-in loss function and default hyper-parameters; (2) **Selected** λ, which applies a custom loss with a different λ for per sensor to minimise the average error across all days; and (3) **Naive baseline**, a simple model that predicts the training mean for each day. The customised loss XGBoost achieves the lowest RMSE at nearly every site, whereas the XGBoost with the default loss function model outperforms the naive baseline, indicating that different λ for each time window can accommodate the spatial sparsity of our network.

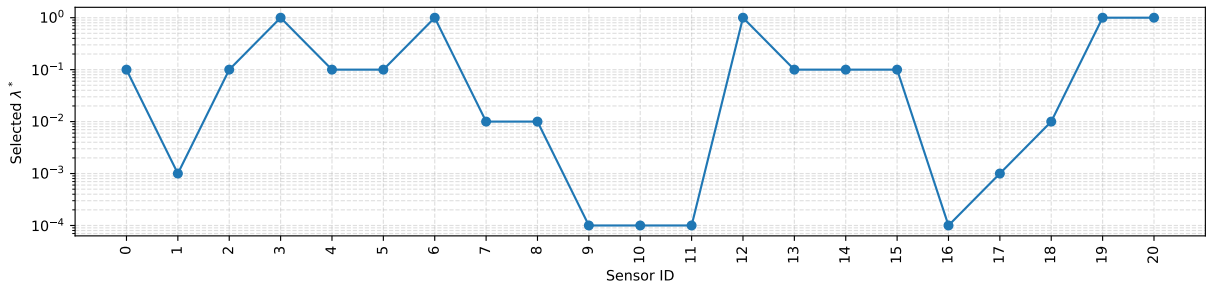


Figure 6.11: Optimal value of the spatial regularisation parameter λ^* for each sensor location in the fitted model (e.g. with a graph-Laplacian penalty). Each point corresponds to one sensor, with larger values of λ^* indicating stronger spatial smoothing imposed on predictions at that location.

Figure 6.11 displays the optimal regularisation parameter λ at every sensor location, with values fluctuating from day to day. This variability indicates that tuning a separate λ for each days dataset is more appropriate than using a single global λ. The RMSE results confirm that XGBoost with the customised loss and individual λ values achieves the lowest errors, outperforming both the naive baseline and standard XGBoost with its default loss function.

6.6 Spatio-temporal conformal prediction

Mao et al. (2024) introduce conformal schemes for spatial data, which are global spatial conformal prediction (GSCP) (marginal validity when locations are random), local spatial conformal prediction (LSCP) (local conditional validity via nearest neighbours), and a smoothed LSCP using spatial kernel weights $\omega_i \propto \exp(-\|s_i - s^\star\|^2/(2\eta^2))$. We extend this to spatio-temporal by introducing a product kernel in space and time, using studentised scores with local bias correction and using effective-sample blending to avoid intervals when N_{eff} is small. To quantify uncertainty in our fine-resolution soil-moisture prediction maps, we wrap our predictor $\widehat{f}$ (XGBoost Δ VWC model) in a conformal prediction framework. Validity relies on the calibration scores, not on the predictor.

Ordinary conformal prediction assumes that residuals are exchangeable. Still, soil moisture exhibits strong spatial and temporal dependence, so past errors from faraway locations or long-ago days can not represent today's uncertainty. We introduce a smoothed, spatio-temporal weighting kernel over the calibration residuals, smoothing them by both spatial distance (via a Gaussian bandwidth h_s) and temporal lag (via exponential decay h_t). After out-of-sample tuning of the parameters $(k_{\text{calib}}, h_s, h_t)$ on calibration days (Lee et al., 2025), this procedure yields prediction intervals

$$\left[\,\hat{y}_j \pm q_j\,\right]$$

that adapt locally to heteroscedasticity in space and time, and guarantee 80 % coverage on average on future days under approximate local exchangeability assumption.

Algorithm 1 Spatio-temporal weighted quantile

Require: Calibration nonconformity scores $\{r_i\}_{i \in \mathscr{I}_{\text{calib}}}$,
locations $\{s_i, t_i\}$ in calib, test point (s^*, t^*), miscoverage α,
spatial bandwidth h_s, temporal bandwidth h_t
Ensure: Weighted quantile $\hat{q}$ at (s^*, t^*)
$m \leftarrow |\mathscr{I}_{\text{calib}}|$
for $i = 1, \ldots, m$ **do**
  $d_i \leftarrow \|s_i - s^*\|$ ▷ Euclidean distance (m)
  $\Delta t_i \leftarrow |t_i - t^*|$ ▷ Time lag (days)
  $w_i \leftarrow \exp\left(-d_i^2/(2h_s^2)\right) \exp\left(-\Delta t_i/h_t\right)$
end for
Sort the pairs $\{(r_i, w_i)\}$ by increasing r_i, yielding $\{(r_{(1)}, w_{(1)}), \ldots, (r_{(m)}, w_{(m)})\}$
$W_{\text{tot}} \leftarrow \sum_{j=1}^{m} w_{(j)}, \quad C \leftarrow 0$
for $j = 1, \ldots, m$ **do**
  $C \leftarrow C + w_{(j)}$
  if $C \geq (1-\alpha) W_{\text{tot}}$ **then**
    $\hat{q} \leftarrow r_{(j)}$
    break
  end if
end for

Algorithm 2 Spatio-temporal conformal prediction

Require: Full dataset $\{(X_i, y_i, s_i, t_i)\}_{i=1}^{N}$, trained model f,
test inputs $\{(X_j^*, s_j^*, t_j^*)\}_{j \in \mathscr{I}_{\text{test}}}$,
miscoverage level α, calibration window k_{calib},
spatial bandwidth h_s, temporal bandwidth h_t
Ensure: Prediction intervals $\left[\hat{y}_j^* \pm q_j^*\right]$ for each $j \in \mathscr{I}_{\text{test}}$
$\mathscr{I}_{\text{train}} \leftarrow \{\, i : t_i < t^* - k_{\text{calib}} \,\}$
$\mathscr{I}_{\text{calib}} \leftarrow \{\, i : t^* - k_{\text{calib}} \leq t_i < t^* \,\}$
Fit f on $\{(X_i, y_i)\}_{i \in \mathscr{I}_{\text{train}}}$
Compute calibration residuals

$$r_i \;=\; \left| y_i - f(X_i) \right| \quad \forall i \in \mathscr{I}_{\text{calib}}$$

for all $j \in \mathscr{I}_{\text{test}}$ **do**
  $\hat{y}_j^* \leftarrow f\left(X_j^*\right)$
  Compute q_j^* by calling **Algorithm 1** at (s_j^*, t_j^*)
  Output interval $\left[\hat{y}_j^* - q_j^*,\ \hat{y}_j^* + q_j^*\right]$
end for

6.6.1 Spatio-temporal smoothed conformal prediction (stLSCP)

We extend LSCP (Mao et al., 2024) to the spatio-temporal setting and propose a method spatio-temporal smoothed conformal prediction (ST-LSCP).

Let $S \in D \subset \mathbb{R}^d$ (d=2) denote space and $T \in \mathbb{R}$ be time. We observe

$$Z_i = (S_i, T_i, X_i, Y_i) = (S_i, T_i, X(S_i, T_i), Y(S_i, T_i)), \quad i = 1, \ldots, n,$$

with features X and response variable Y (volumetric water content, VWC). We fit an XGBoost predictor $\widehat{f}$ for the next-day change $\Delta Y(S,T) = Y(S,T+1) - Y(S,T)$ and produce point prediction at a target location and time $(s^\star, t^\star)$:

$$\widehat{y}(s^\star, t^\star) \;=\; Y(s^\star, t^\star - 1) \;+\; \widehat{f}\big(X(s^\star, t^\star - 1)\big).$$

Spatio-temporal localisation (weights).
To ensure that calibration scores are most relevant for prediction at a given spatio-temporal location $(s^\star, t^\star)$, we localise the conformal procedure by re-weighting the calibration residuals according to the spatio-temporal proximity. To be specific, rather than treating all calibration residuals equally, we assign higher weights to calibration points that are close in both space and time to test point $(s^\star, t^\star)$. This is done with a product kernel in space and time:

$$\tilde{\omega}_i(s^\star, t^\star) \;\propto\; \exp\!\left(-\frac{\|s_i - s^\star\|^2}{2h_s^2}\right) \exp\!\left(-\frac{|t_i - t^\star|}{2h_t^2}\right),$$

Let $\tilde{\omega}_j$ denote the unnormalised weight. We normalise by

$$\omega_j = \frac{\tilde{\omega}_j}{\sum_{i \in n} \tilde{\omega}_i}, \qquad \sum_{j \in n} \omega_j = 1.$$

where h_s and h_t are spatial and temporal bandwidths; n is the number of calibration observations (nonconformity scores) before the test time point, which is the location and time point pair in the calibration window. In this way, the calibration scores are effectively adapted to local conditions, making the prediction intervals reflect the spatio-temporal condition of the test point. When $h_t \to \infty$, it recovers spatial LSCP. $h_s, h_t \to \infty$ recovers global spatial CP (GSCP) (Mao et al., 2024).

Non-conformity (studentised residuals) and weighted quantile.
After defining how calibration points are weighted, the next step is to compute the non-conformity scores, which measure predictive error. On a calibration window before test point t^*, we compute signed residuals for the change (Lei et al., 2018),

$$R_i^{\text{sign}} = \Delta Y_i - \widehat{f}(X_i), \qquad R_i = |R_i^{\text{sign}}|.$$

Since raw residuals can vary in scale across space and time, we stabilise them by robust studentisation. To be specific, we studentise via a median absolute deviation (MAD) (Rousseeuw and Croux, 1993), which is a robust measure of spread at (s^*, t^*):

$$b^*(s^*, t^*) = \text{wMed}\{\, R_i^{\text{sign}},\ \omega_i(s^*, t^*) \,\}$$

$$\mathrm{MAD}^*(s^*,t^*) = \mathrm{wMed}\{\,|R_i^{\mathrm{sign}} - b^*(s^*,t^*)|,\ \omega_i(s^*,t^*)\,\},$$

where the $b^\star$ estimates the systematic local bias of the predictor around (s^*,t^*) and define studentised scores $\widetilde{R}_i(s^*,t^*) = \frac{|R_i^{\mathrm{sign}} - b^*(s^*,t^*)|}{\mathrm{MAD}^*(s^*,t^*)+\varepsilon}, \quad \varepsilon > 0$ with a small $\varepsilon > 0$.

The weighted empirical CDF and $(1-\alpha)$ quantile at $(s^\star,t^\star)$ are

$$\widehat{F}_n(r \mid s^\star,t^\star) = \sum_{i=1}^{n} \omega_i^{(n)}(s^\star,t^\star)\,\mathbf{1}\{\widetilde{R}_i \le r\}, \quad \widehat{q}_{1-\alpha}(s^\star,t^\star) = \inf\{r : \widehat{F}_n(r \mid s^\star,t^\star) \ge 1-\alpha\}.$$

We convert it back to the original scale via $q^\star = \widehat{q}_{1-\alpha} \cdot \mathrm{MAD}^*(s^*,t^*)$.

Interval construction.
From the localised half-width $q^\star$, we can now form the prediction interval at $(s^\star,t^\star)$.

$$y^{\mathrm{bc}}(s^\star,t^\star) = y(s^\star,t^\star) + b^*(s^*,t^*)$$

$$\Gamma_\alpha^{\mathrm{ST}}(s^\star,t^\star) = \left[\,\widehat{y}^{\mathrm{bc}}(s^\star,t^\star) - q^\star,\ \widehat{y}^{\mathrm{bc}}(s^\star,t^\star) + q^\star\,\right].$$

Training and calibration steps.
For each test day $t^\star$:

1. **Training set:** all samples with $T \le t^\star - k_{\mathrm{cal}} - 1$. Fit $\widehat{f}$ (we use XGBoost on ΔY).
2. **Calibration set:** the k_{cal} days immediately before $t^\star$: $t^\star - k_{\mathrm{cal}}, \ldots, t^\star - 1$.
3. Compute residuals R_i (and $\widetilde{R}_i$) on the calibration set only and build $\Gamma_\alpha^{\mathrm{ST}}(s^\star,t^\star)$ using the weighted quantile above.

Assumptions.
As for the theoretical properties. The spatio-only GSCP/LSCP theory shows: (i) with randomly sampled locations, GSCP has finite-sample marginal coverage; (ii) under spatial infill (locations become dense near $s^\star$), LSCP has asymptotic conditional coverage at $s^\star$ (Mao et al., 2024). Our stLSCP extends this to spatio-temporal data. Under a spatio-temporal infill regime (the spatial domain $D \subset \mathbb{R}^d$ and time interval $T \subset \mathbb{R}$ are fixed while sampling becomes increasingly dense in $D \times T$) on $D \times T$, with $h_s \to 0$, $h_t \to 0$ and $n h_s^d h_t \to \infty$. The calibration scores with non-negligible weight concentrate in a vanishing neighbourhood of $(s^\star,t^\star)$ and become approximately exchangeable with the test score. Thus, the weighted empirical quantile $\widehat{q}_{1-\alpha}(s^\star,t^\star)$ consistently estimates the conditional $(1-\alpha)$ quantile of the test score, yielding asymptotic conditional validity at

$(s^\star, t^\star)$.

It is noted that the spatio-temporal weighting breaks the strict exchangeability assumption, so the resulting intervals do not have finite-sample marginal coverage guarantees. Instead, we validated calibration empirically by an expanded window over multiple days, selecting $(k_{\text{calib}}, h_s, h_t, \alpha)$ to target $1-\alpha = 0.80$ and obtaining mean coverage ≈ 0.80 on held-out days. These hyperparameters were then frozen before applying to the test day.

Effective-sample blending

While the asymptotic coverage holds for the spatio-temporal scenario, in practice, our calibration windows are finite. Residuals are dependent, and we assume approximate local exchangeability of residuals within each calibration window. To investigate instability, we use the effective sample blending. If the effective number of calibration points is too small, local intervals may be unreliable. We compute $N_{\text{eff}} = \frac{1}{\Sigma_{i=1}^{n}\left(w_i\right)^2}$. and blend local and global half-widths via $\tau = \frac{N_{\text{eff}}}{N_{\text{eff}}+N_0}$ and $q_{\text{final}} = \tau\, q_{\text{local}} + (1-\tau)\, q_{\text{global}}$, where q_{global} is the studentised $(1-\alpha)$ quantile with equal weights for all the points (Quiñonero-Candela et al., 2022; Reddi et al., 2015).

Hyperparameter selection.

Finally, the method needs to choose the bandwidths (h_s, h_t), calibration window length k_{cal}, and blending constant N_0. We select these hyperparameters by rolling temporal cross-validation across multiple days. For each candidate window, we compute empirical coverage and average the interval width. The chosen setting minimises the coverage gap $|\widehat{\text{cov}} - (1-\alpha)|$, maintains reasonable width and enforces a floor on N_{eff}, which prevents calibration being driven by a few highly weighted points. This ensures that the resulting intervals are both valid on average and not very wide.

Conformal prediction (CP) is model-agnostic, and the key assumption is exchangeability. In spatio-temporal data, we approximate local exchangeability within a neighbourhood of (s^*, t^*) (for approximate exchangeability). This makes CP suitable even when error scales vary across space and time. The soil moisture data exhibit both strong spatial heterogeneity (different sensor locations have different error distributions) and temporal non-stationarity (error scales change over time), so we use two complementary conformal prediction settings to provide conformal coverage both within a day (across sensors on a fixed day) and between days (across successive days at a fixed location):

- **Temporal conformal calibration** Provides marginal validity over time across all locations. For any test day T, we hold out the k days immediately before T to calibrate a global half-width q, ensuring that across many days our intervals cover $100(1-\alpha)\%$ of observations in time.

- **Per-sensor conformal prediction** Provides conditional validity at each fixed sensor location. We leave each sensor out in turn (LOSO) to tune and calibrate intervals that guarantee roughly $1-\alpha$ coverage at that particular site on the same day.

We compare these two validation strategies: the former ensures overall temporal coverage across the entire grid, while the latter gives stronger local guarantees at sensor sites.

6.6.2 Temporal conformal calibration

To account for temporal dependence in VWC prediction, we use a rolling split-conformal scheme over the last k days. For each test day T:

1. **Select test day T.** We set uncertainty bands for VWC on day T.

2. **Define calibration window.**

   The k days immediately before T define the calibration set.

   $$\mathscr{C}_T = \{T-k,\, T-k+1,\, \ldots,\, T-1\}$$

3. **Train on earlier days.**

   Fit the spatiotemporal predictor $\hat{f}$ using only data from days $\leq T-k-1$. Let $\hat{y}_i$ denote its predictions on $\mathscr{C}_T$ and y_i the truths.

4. **Compute calibration residuals.**

   $$r_i = \left| y_i - \hat{y}_i \right| \quad \text{for } i \in \mathscr{C}_T.$$

5. **Choose the conformal half-width.**

   Global (unweighted) split-CP:

   $$q = \text{Quantile}_{1-\alpha}\left\{ r_i :\ i \in \mathscr{C}_T \right\}.$$

   Weighted (to anticipate stLSCP):

   Assign proximity weights $w_i(T)$ (e.g. $w_i(T) \propto \exp(-|t_i - T|/h_t)$ for temporal or $w_i(s^\star, T) \propto \exp(-\|s_i - s^\star\|^2/2h_s^2)\exp(-|t_i - T|/h_t)$ for spatio-temporal), normalise $p_i = w_i/\sum_j w_j$, and take the weighted $(1-\alpha)$ quantile q.

6. **Predict with intervals on day T.**

   For each location, form the point predict $\hat{y}_{\text{new}}$ and the interval

   $$\left[\hat{y}_{\text{new}} - q\ ,\ \hat{y}_{\text{new}} + q \right].$$

Bias-corrected version used in stLSCP: Define the local bias

$$b^\star \;=\; \mathrm{wMed}\{\, y_i - \hat{y}_i \;,\; w_i(s^\star, T)\},$$

and the bias-corrected centre

$$\widehat{y}^{\mathrm{bc}}(s^\star, T) \;=\; \widehat{y}(s^\star, T) \;+\; b^\star,$$

then report

$$\big[\widehat{y}^{\mathrm{bc}}(s^\star, T) - q \;,\; \widehat{y}^{\mathrm{bc}}(s^\star, T) + q\big].$$

Rolling split-CP yields exact finite-sample marginal coverage $1-\alpha$ under exchangeability between C_T and day T. With temporal dependence, we validate coverage empirically through back testing. The window length k is tuned by rolling CV to minimise the coverage gap and, secondarily, mean interval width.

This step guarantees, under mild exchangeability in time, across many days, the intervals will cover the true VWC approximately $100(1-\alpha)\%$ of the time. In spatio-temporal extension, we give a weight to each residual based on its spatial and temporal proximity to each test point.

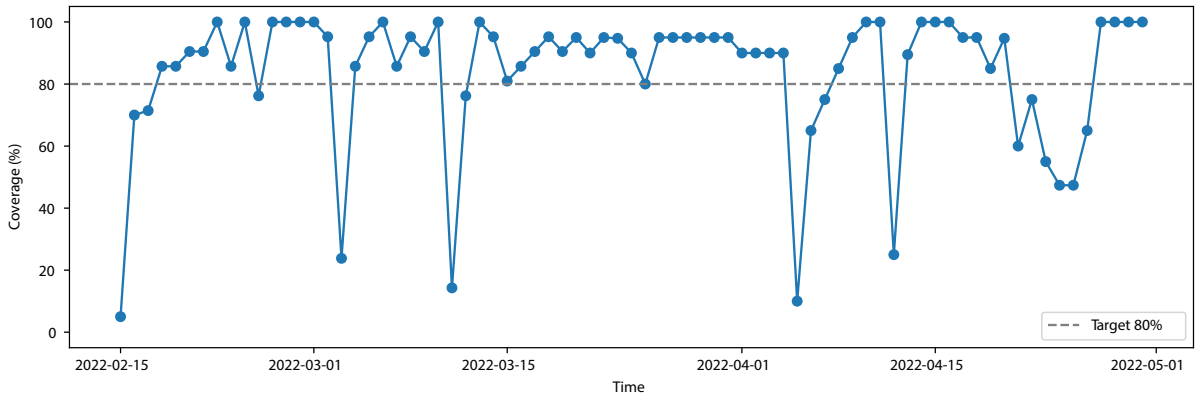


Figure 6.12: Temporal cross-validation coverage. Daily coverage across sensors for each test day using a 30-day rolling training window, 14-day calibration window, and spatio-temporal weighted conformal intervals (h_s=5 km, h_t=12 d) with different τ selected for each training window. The dots represent daily coverage, while the dashed line indicates the 80% coverage target. The averaged coverage = 84.0% (95% CI [82.2%, 85.9%], N =1522); mean interval width = 2.415

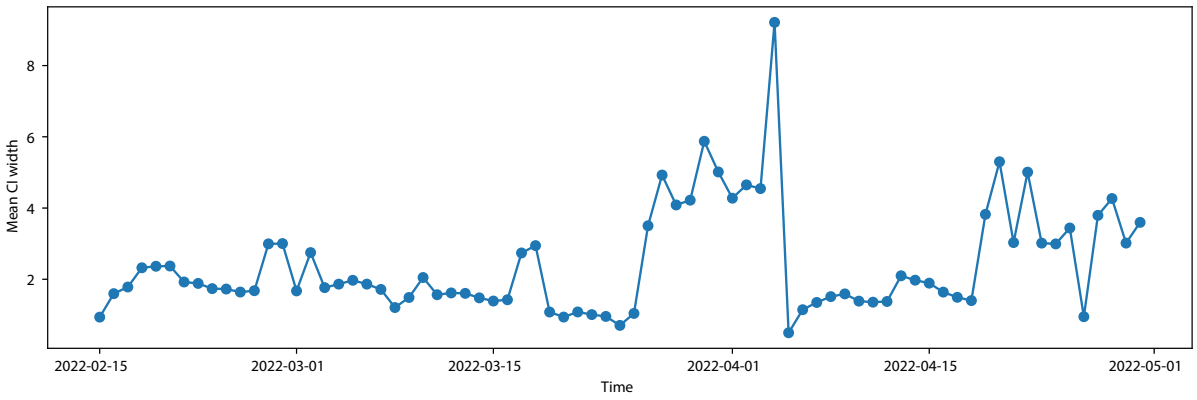


Figure 6.13: Mean interval width. Daily average conformal interval width across sensors for each test day, using a 30-day rolling training window, 14-day calibration window, and spatio-temporal weighted conformal intervals (h_s=5 km, h_t=12 d).

Figure 6.12 demonstrates that daily coverage is usually above 80% target, with a long-term coverage of 84.0% (95% CI [82.2%, 85.9%]) among the 1522 data points. Figure 6.13 shows the mean CI interval width, which explains that on most of the days, the width of the CIs is around 1% to 3%, which matches with the high coverage rates (85% to 100%) in Figure 6.12. When the model detects high local uncertainty or a change in the residual distribution, the CI width expands (e.g., in early April, the width goes up to approximately 9%). The few non-covered test days (e.g., 2022-04-05) occur when the VWC has a rapid change compared to the 14-day calibration window. The sudden increase in the width of the CIs suggests that the spatio-temporal weighting is actually doing its job. It is noted that the width of the CIs and the coverage rates are related, but not in a one-to-one case, because the small number of sensors will impact the coverage rate. The width is calculated based on the calibration data. However, the test day might have a swift change, and the mean width is an average of the sensors. Some sensors that are hard to predict can reduce the coverage rate, even though the mean width appears normal. Therefore, days with a wide width often have high coverage. The wide width with low coverage flags the rapid change or bias that the quantile didn't capture. The narrow width with high coverage means those are easy days.

6.6.3 Per sensor conformal prediction

To obtain conditional validity at each fixed sensor location, we perform a leave-one-sensor-out (LOSO) conformal calibration. In each fold, the entire time series of one sensor is held out for calibration, and a separate interval is constructed for that sensor. Repeating this over all sensors provides per-site guarantees of approximately $100(1-\alpha)\%$ coverage.

1. **Custom loss model at tuned λ** Using the training sensors (all except the heldout one), fit the XGBoost model with the spatial penalty objective $\text{obj}(\cdot;\lambda)$.

2. **Predict on heldout sensor**

$$\hat{y} = f(X_{\text{test}}).$$

3. **Compute absolute residuals**

$$r = \left| y_{\text{true}} - \hat{y} \right|.$$

4. **Calibrate quantile**

$$q = \text{quantile}_{1-\alpha}\{r\} \qquad \text{(e.g. } 80^{\text{th}} \text{ percentile for } \alpha = 0.2\text{)}.$$

5. **Form prediction interval**

$$\left[\, \hat{y} - q,\, \hat{y} + q \,\right],$$

which (empirically) covers $(1-\alpha) \times 100\%$ of held-out sensor values.

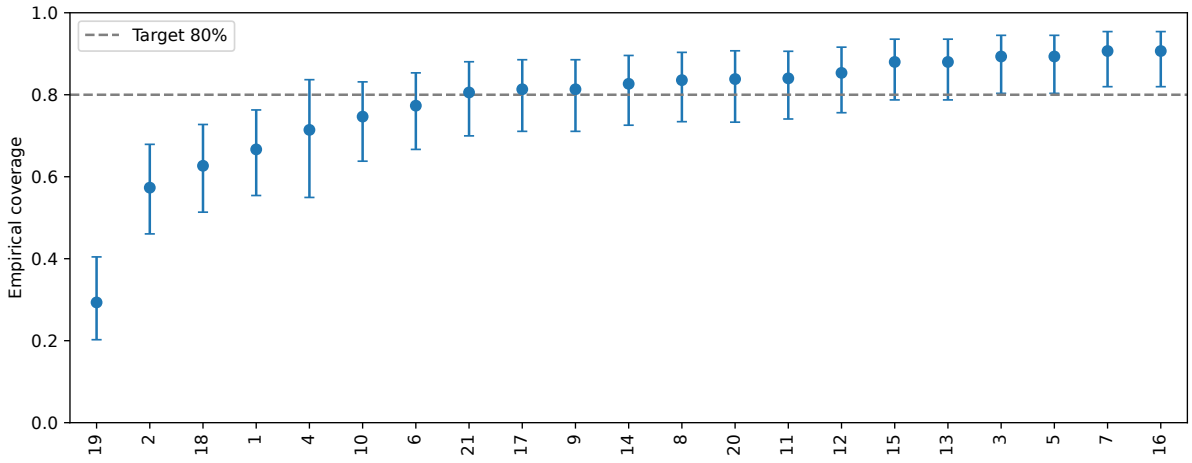


Figure 6.14: Spatial LOSO-CV for each sensor coverage (30-day training window and 14-day calibration window, $\alpha = 0.20$). Each dot is a sensors empirical coverage across all leave-one-out test days. The bars are 95% CIs for that coverage. Sensors are sorted left-to-right by coverage. The dashed line marks the 80% target (1-α).

Figure 6.14 demonstrates the percentage of each sensors test days when the true VWC fell into the conformal interval when that sensor was held out of training. The error bar reflects sampling uncertainty from the number of test days for each sensor. If the entire CI is below 0.80, the prediction of that sensor is not reliable. Points clearly above 0.80 indicate a reliable prediction at that site. The low-coverage sensors usually match high local variability, typically on the edge or in very sparse areas.

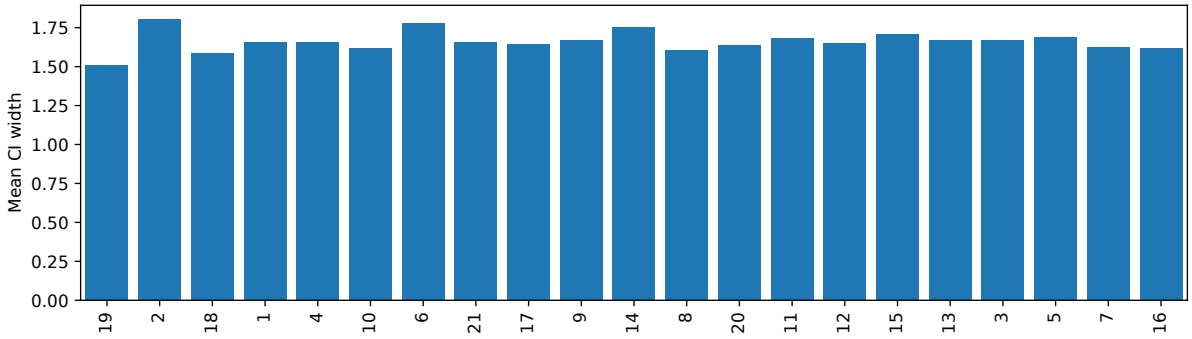


Figure 6.15: Mean conformal CI width for each sensor under LOSO-CV (30-day training window, 14-day calibration window, $\alpha = 0.20$).

Figure 6.15 demonstrates that the wider bars indicate locations where the method gives higher uncertainty, which suggests either higher local variability or poor support from nearby calibration residuals.

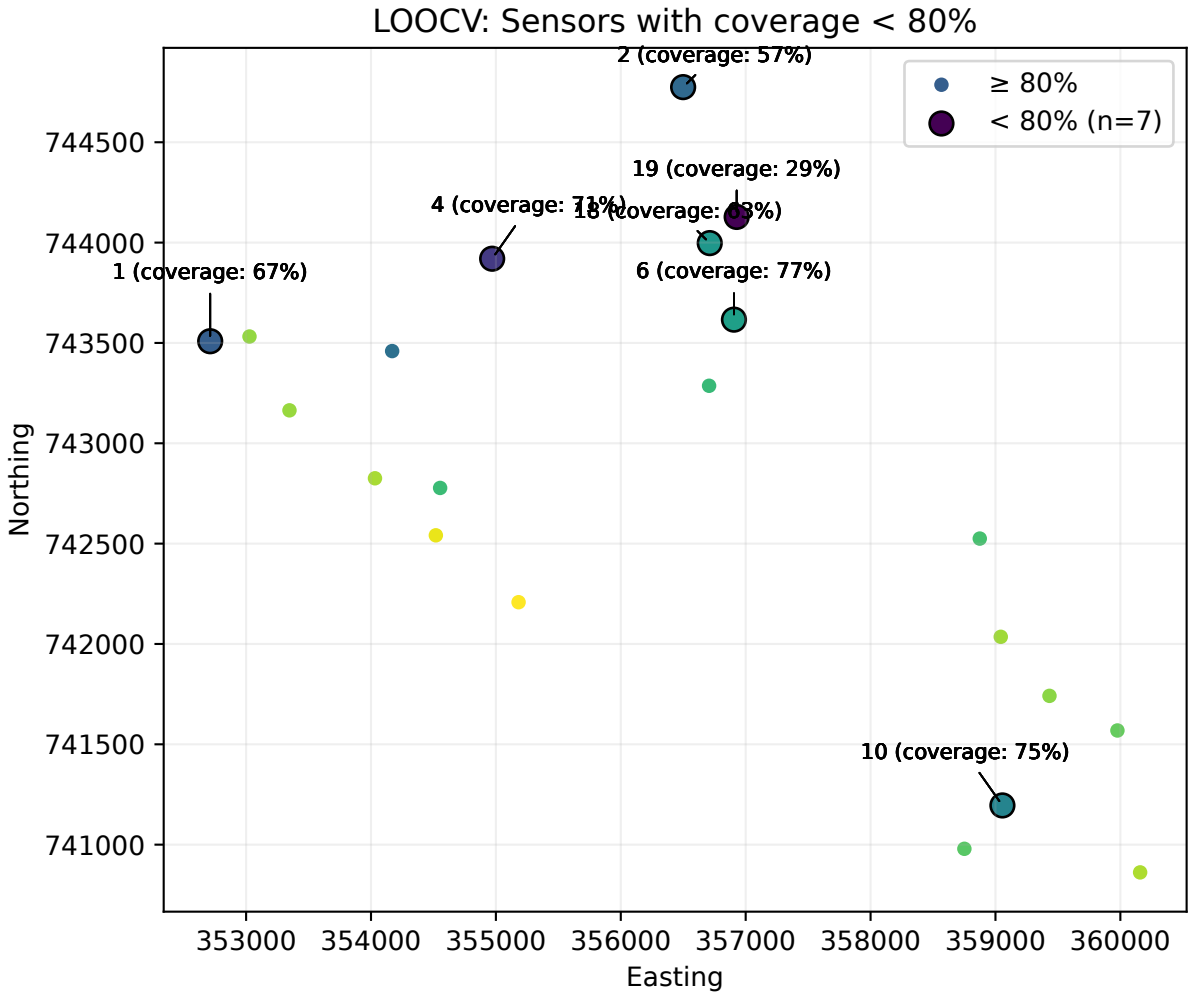


Figure 6.16: Distribution of all sensors under LOSO-CV using a 30-day training window and 14-day calibration window ($\alpha = 0.20$). Circles show sensor locations, and the background points are sensors meeting the 80% target.

Figure 6.16 shows patterns for the low-coverage sensors. There is a spatial cluster, several edge-of-domain sites, locations with very few nearby neighbours, and a few isolated under-coverage points. These patterns suggest that the low coverage may be driven by boundary effects, sparse local support, sensors close to the rivers, or missing covariates, rather than random error alone.

6.6.4 Spatio-temporal smoothed conformal prediction

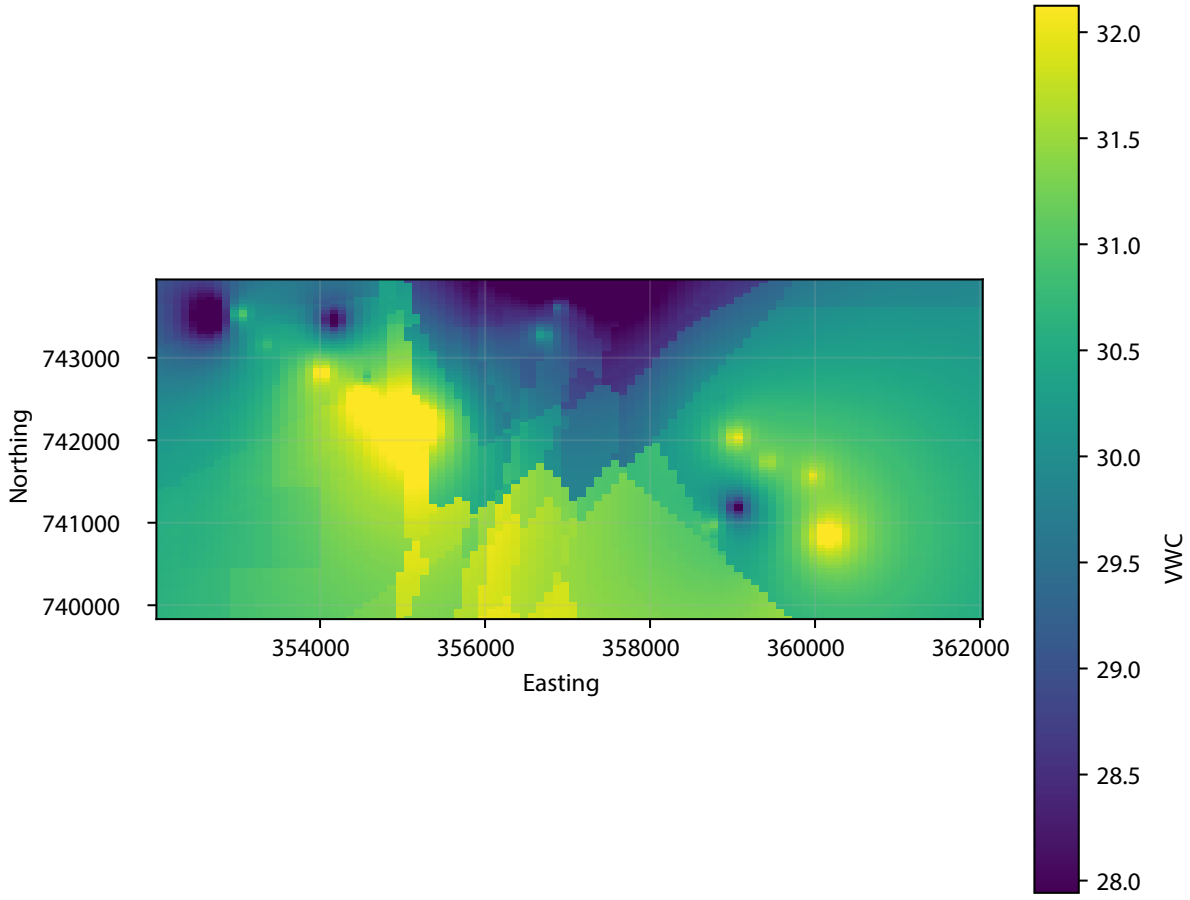


Figure 6.17: Fine-resolution (100 × 100) prediction of near-surface volumetric water content (VWC) across the study catchment for 1 April 2022, obtained from the fitted spatio-temporal model. The map highlights spatial variability in soil moisture at a fine spatial scale, with higher predicted VWC corresponding to wetter soils and lower values indicating drier areas.

Figure 6.17 demonstrates the 100 × 100 prediction map of VWC for 1 April 2022 across the study catchment. The broad gradients across the space suggest large-scale covariates such as elevation and SWI, while smaller patterns around some areas suggest a stronger influence from nearby sensors via the KNN weighting. Subtle stripe patterns may suggest where coarse grid covariates change or where the K-nearest-neighbour set switches. These are input impacts rather than physical discontinuities.

6.7 Comparison between the INLA-SPDE and the XGBoost conformal prediction

We compare one-day-ahead VWC predictions from a Bayesian INLA-SPDE data fusion model (Chapter 5) and a gradient boosted tree ensemble (XGBoost) calibrated with spatio-temporal locally weighted conformal prediction (Chapter 6). Figure 6.18 shows LOSO-CV RMSE on 2022-04-01. XGBoost obtains lower RMSE because it flexibly learns non-linear effects and interactions among features (including previous-day VWC and gridded covariates), with little shrinkage. INLA-SPDE estimates a latent Gaussian field with a Matérn covariance matrix and fuses point and gridded data. This includes spatial smoothing, which is advantageous in sparse regions, but can slightly smooth out local changes, thereby increasing point error in heterogeneous areas. Figure 6.19 compares uncertainty between the predictions from these two models. INLA-SPDE intervals are Bayesian posterior predictive intervals that combine latent process, measurement, and parameter uncertainty (and handle covariate misalignment), and are therefore often broader. Conformal intervals around the XGBoost prediction are distribution-free predictive intervals with nominal marginal coverage. The spatio-temporal weighting adapts the widths to local data support, which narrows near dense areas, recent information, and widens when support is weak. In summary, XGBoost with CP focuses on empirical point accuracy with calibrated, locally adaptive coverage, whereas INLA-SPDE provides a Bayesian hierarchical data fusion framework whose accuracy depends on the hierarchical assumptions and priors, and which can outperform in sparse regions and smoothly different regions.

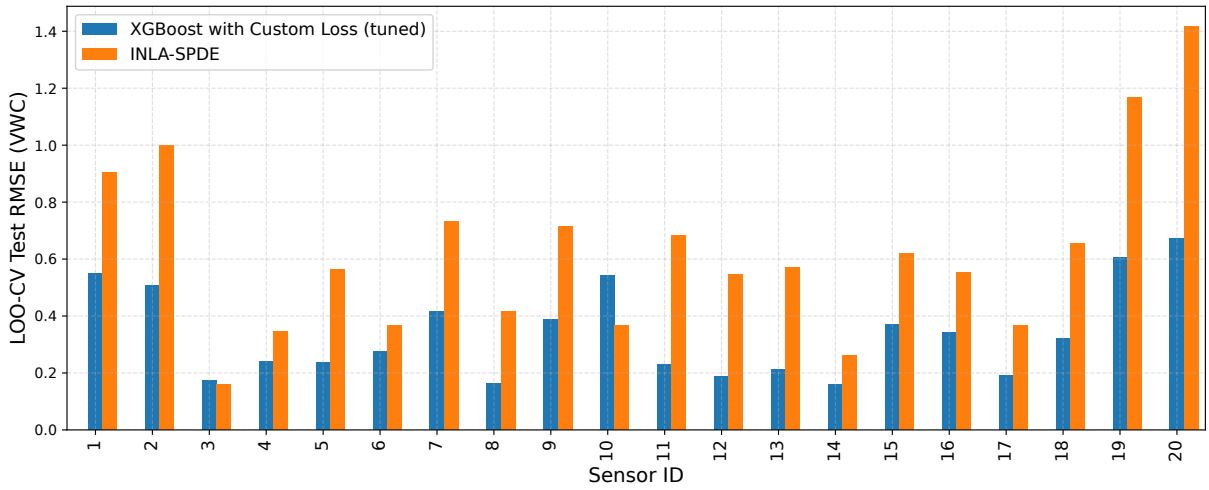


Figure 6.18: Root mean squared error (RMSE) from leave-one-sensor-out cross-validation (LOSO-CV) comparing a Bayesian hierarchical INLA-SPDE model and a gradient-boosted tree ensemble (XGBoost) for one-day-ahead volumetric water content (VWC) prediction on 2022-04-01. Lower RMSE values indicate better predictive performance.

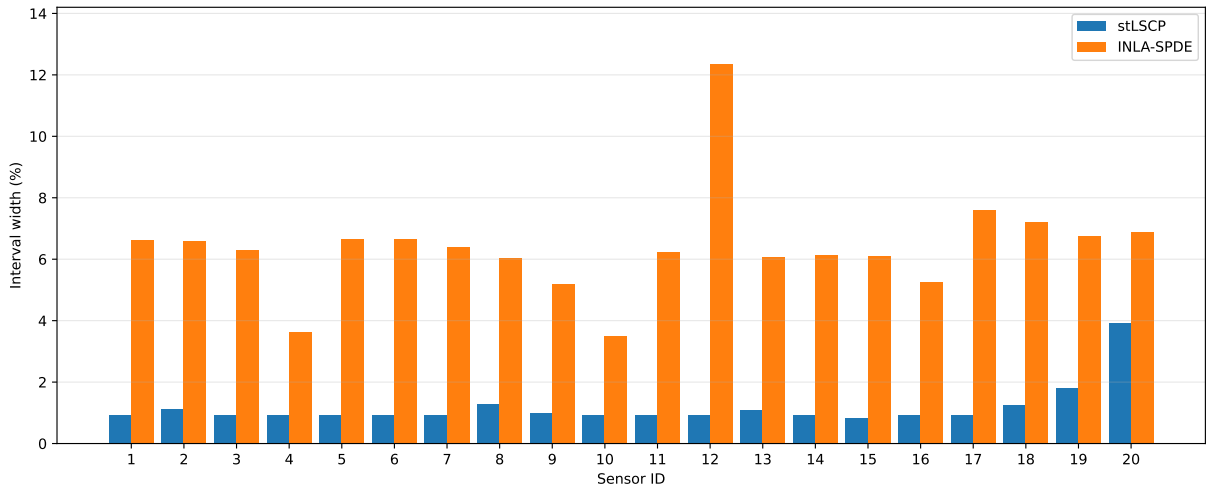


Figure 6.19: Interval widths ($\alpha = 0.2$) for Conformal Prediction (CP) and INLA-SPDE by sensor for 2022-04-01. CP widths are derived from next-day, locally calibrated conformal residuals, while INLA-SPDE widths come from the spatial models predictive intervals computed from the posterior.

6.8 Conclusion

This chapter introduces a spatio-temporally weighted conformal prediction framework (stLSCP) built on an XGBoost data fusion model that combines point data and gridded data. For the XGBoost data fusion model, we modify XGBoost by introducing a graph Laplacian penalty term controlled by λ in a custom loss function to encourage spatial and temporal smoothness. We tune λ via cross-validation across multiple time points and leave-one-sensor-out cross-validation (LOSO-CV). The custom loss function shows consistent improvement from both spatial and temporal perspectives compared to the default loss function.

The spatio-temporally weighted conformal prediction framework (stLSCP) is used to quantify the uncertainty of the XGBoost data fusion model. To overcome the avoidance of the exchangeability assumption in ordinary conformal prediction, we introduce a smoothed, spatio-temporal weighting kernel over the calibration residuals, which down-weights the residuals by both spatial distance (via a Gaussian bandwidth h_s) and temporal lag (via an exponential decay h_t). The stLSCP is also being validated using spatial and temporal cross-validation.

One-day-ahead VWC predictions performance is compared between a Bayesian INLA-SPDE data fusion model (Chapter 5) and a gradient boosted tree ensemble (XGBoost) calibrated with spatio-temporal locally weighted conformal prediction (Chapter 6). In LOSO-CV on 1 April 2022, XGBoost obtains a lower RMSE than the Bayesian INLA-SPDE model. For uncertainty, INLA-SPDE obtains posterior predictive intervals which rely on the model assumptions. Our conformal intervals obtain marginal coverage and are adapted to local support, although they are

predictive and can be wider in regions of low support. Overall, the proposed method offers a practical way to accurately predict a next-day soil moisture map calibrated with locally adaptive uncertainty, which complements INLA-SPDE's fully probabilistic model under parametric assumptions.

In summary, this chapter develops a spatio-temporally constrained ensemble data fusion model with conformal prediction, which integrates spatio-temporal structure to connect modern ML with established statistical modelling. We include a comparison between a Bayesian hierarchical data fusion model (INLA-SPDE) and XGBoost with conformal prediction. Since the main aim of this work is to present a novel data fusion model and compare the performance between two different model structures, the investigation of how these two models perform on multiple days is left for future work. Thus, this should not be seen as a general measure of goodness-of-fit, as performance varies by day. It is noted that the model structure is different: XGBoost conditions directly on previous-day point data (via KNN), whereas INLA-SPDE incorporates it through a continuous latent field, which typically produces wider predictive intervals.

Chapter 7

Conclusions and discussion

This thesis introduces three INLA-based data fusion methods: spatio-temporal regression with misaligned covariates, a spatial data fusion method, and a spatio-temporal data fusion method, together with an XGBoost based constrained ensemble method with conformal prediction. These all developed to merge in-situ point and satellite gridded data under different spatio-temporal supports. The research question is motivated by the in-situ soil moisture data provided by SEPA in Elliot Water and the satellite images provided by Copernicus. It is necessary to develop a data fusion method of point data and gridded data, so that the accuracy of the in-situ data can be combined with spatial and temporal information from satellite data to generate a fine-resolution map with uncertainty quantification.

The literature review discusses the methodology from the existing data fusion literature that is relevant to my thesis. Statistical data fusion method mainly follows two mainstream frameworks: Bayesian hierarchical models (BHM) and machine learning (ML). BHMs provide uncertainty propagation and explicit spatio-temporal structure, but they can be computationally expensive and sensitive to prior and model structural choices. ML scales well and can capture nonlinear relationships, but often treats observations as independent and lacks uncertainty quantification. Although there are many existing data fusion studies, there are several gaps that need to be addressed: Firstly, the change of support between point sensors and satellite grids is not always handled properly. Secondly, spatial and temporal dependence are not included in the ML framework. Finally, uncertainty quantification is usually missing. This thesis addresses these gaps by developing and comparing a geostatistical (INLA-SPDE) model and a spatially regularised ML (XGBoost) data fusion framework that combines in-situ and satellite data to generate high-resolution soil moisture maps and one-day-ahead predictions with uncertainty quantification. Taken together, the methods and case studies in this thesis provide frameworks for fusing spatially misaligned point and gridded data in environmental applications. Beyond the specific Elliot Water case study, the work shows how INLA-SPDE models and spatially regularised XGBoost can be used in a complementary way to combine in-situ sensors with satellite products, handle change of support, and deliver high-resolution predictions with calibrated uncertainty. The simulation

studies quantify how sensor density, grid resolution and missingness influence performance, offering practical guidance for the design of monitoring networks. These advances are directly relevant to soil moisture mapping but can also be transferred to other environmental variables where multiple imperfect data sources must be combined.

7.1 Exploratory data analysis

Chapter 2 explores in-situ volumetric water content (VWC), the satellite data soil water index (SWI), and COSMOS data to characterise spatio-temporal patterns in soil moisture and other relevant covariates. After data preprocessing, the exploratory analysis shows a strong seasonal pattern and temporal autocorrelation. For example, VWC changes positively with recent precipitation and negatively with air temperature. Spatial autocorrelation indicates that neighbour locations share similar hydrological behaviour, while temporal autocorrelation shows long-term and short-term memory in soil moisture, which suggests that it is necessary to consider both spatial and temporal autocorrelation.

The relationship between point VWC and gridded SWI is similar but nonstationary. Pearson correlations (0.61 to 0.72 across sites) suggest a moderate linear relationship, and the rolling (15-day) correlations suggest time-varying dependence that is strong during wet periods and weak during dry periods. Cross-correlation analysis suggests that VWC leads SWI by roughly 2 to 3 days at several locations, which gives insights into the data fusion model.

These findings motivate several key points for the data fusion modelling. First, the change of support must be handled to connect the point and gridded processes. Second, models should capture persistence and seasonality via low-order autoregressive components and covariates, while allowing nonlinear effects and interactions. Third, spatial structure should be included to consider neighbourhood information. Finally, because the data fusion model incorporates multiple data sources, uncertainty quantification is important for reliable inference and prediction.

In summary, the EDA demonstrates characteristics of in-situ, COSMOS, and satellite datasets. These characteristics directly inspire the two main model frameworks in the thesis: a spatio-temporal INLA-SPDE model and a spatially regularised machine-learning prediction with locally calibrated uncertainty. Both of them are designed to introduce spatio-temporal dependence into the data fusion model and deliver high-resolution soil moisture predictions with uncertainty quantification.

7.2 Spatio-temporal regression with misaligned covariates

Part of this work (Spatial regression with misaligned covariates for soil moisture mapping) is published in the Proceedings of the 38th International Workshop on Statistical Modelling (IWSM), Durham, UK, 14-19 July 2024.

Chapter 3 develops a spatial regression framework for soil moisture mapping that incorporates misaligned covariates within the INLA-SPDE data fusion framework. The model links rainfall (misaligned), soil temperature (aligned), and VWC (response) through Matérn Gaussian random fields. VWC is modelled as a function of a fixed effect elevation covariate and on scaled copies of the latent spatial effects from rainfall and temperature. Simulation study shows that fixed effects are estimated well, whereas range parameters and some scaling coefficients are sometimes weakly identified, and both coverage and accuracy improve as the number of locations increases.

A purely spatial data fusion model in Eq.(3.6) is compared with its spatio-temporal extension in Eq.(3.7). This study shows that introducing multiple time points improves inference and prediction at unknown locations: moving from a single time point ($k = 1$) to multiple time points (e.g., $k = 30$) generates narrower intervals and reduced bias, demonstrating how temporal information can compensate for sparse spatial coverage by borrowing information over time. In the real-data application, however, the predicted mean surface shows limited spatial variation away from sensors. This is likely because in areas without sensors, the model relies on covariates like elevation, so elevation dominates the spatial pattern.

To increase spatial support, the gridded satellite data are incorporated in a joint change-of-support framework (INLA-SPDE fusion of point and gridded data) in the next chapter.

7.3 Data fusion method for the spatial-only model

Chapter 4 evaluates a spatial-only INLA-SPDE data fusion framework for combining point and gridded data, using two simulation studies and a real-data application. Firstly, the spatial-only INLA-SPDE data fusion model is extended to a spatio-temporal model that fuses point and gridded data by mapping observations to a Gaussian Markov Random Field (GMRF) via a novel projection matrix (Moraga et al., 2017). Secondly, it shows how different factors impact parameter estimation and model prediction. Across different latent-field smoothness levels, the point models RMSPE_y is greater than that of the grid and joint models when sensor density is low, but the difference disappears as the number of sensors increases. Thirdly, the joint models advantage becomes more beneficial with dense point data. Performance drops when grid data availability decreases from 80% to 10%, and remains stable above 80%, which shows the model is robust to moderate missingness of the gridded data. Fourthly, the benefit of the joint model depends on grid resolution: at fine grid resolution, gridded data already capture

small-scale structure, so point data adds little contribution. At a coarser grid resolution, point data contributes more information and improves predictions. Moreover, differences in measurement error between point and grid data sources affect the data fusion prediction results, which suggests it is necessary to model error structure explicitly. Finally, mesh construction influences both prediction accuracy and computational cost, and should be selected to balance between model fitting and computational cost. The next chapter extends this data fusion framework to the spatio-temporal setting, borrowing information across multiple time points and incorporating different measurement errors within a spatio-temporal data fusion model.

7.4 Spatio-temporal data fusion model

Chapter 5 generalises the spatial-only framework to a spatio-temporal data fusion model that integrates multiple days of data. The latent process uses Matérn spatial covariance extended in time via an AR(1) temporal dependence structure. A simulation study varying the number of days ($k \in \{3, 7, 10, 30\}$) shows that more temporal points enhance estimation and prediction for $\alpha_1, \alpha_2, \beta_1, \beta_2$. The RMSE($\beta_1$) decreases by 35% from 3 to 30 days. The joint (point and grid) model consistently outperforms point and grid models (at $k = 30$, RMSE reduction of around 15-20% for scaling parameters and around 10-15% for spatial variances) and is more robust under sparse data. However, persistent biases in α_3 and range ρ remain across different k, and while 95% credible intervals narrow with more time points, they do not overcome these structural errors, which suggests modelling choices such as spatio-temporal covariances separability, fixed smoothness, and simple AR(1) temporal dependence structure.

7.5 Spatio-temporally constrained ensemble learning with conformal prediction: A distribution-free approach to uncertainty-aware data fusion

Chapter 6 introduces a spatio-temporally weighted conformal prediction (stLSCP) framework built on an XGBoost point and grid data fusion model with a custom loss function including graph-Laplacian penalty (tuned via cross-validation), which obtains consistent accuracy improvement over the default loss function and provides locally adaptive, distribution-free uncertainty through spatio-temporal residual weighting. Spatial and temporal cross-validation confirmed reliable empirical coverage. In a LOSO-CV comparison, XGBoost with stLSCP attained a lower RMSE than the Bayesian INLA-SPDE model. However, this comparison is just an example rather than a general performance, as performance varies from day to day. It is noted that the models have different model structures, and the XGBoost conditions directly on previous day point data (via KNN), whereas INLA-SPDE propagates information through a continuous latent field and an AR(1) process in Eq.(5.1), so INLA-SPDE produces broader, spatially smooth predictive

intervals that arise from partial pooling via the latent field. To be specific, the latent Gaussian field has spatial correlation, so each sites estimate is partially pooled with nearby sites (via Matérn covariance structure). This leads to spatially coherent estimates and wider but well calibrated intervals in sparse areas, while the purely local methods may have narrower but uneven intervals (sharp jumps rather than smooth change).

7.6 Discussion, limitations and future work

Based on the previous chapters results and limitations, several things are identified for future work. It is noted that different temporal supports can be handled inherently in the spatio-temporal INLA-SPDE data fusion model. However, the in-situ sensor data are averaged from 15-minute to daily resolution to align with the satellite gridded data. Future research may investigate how changes in temporal support affect model performance, but the main aim of this thesis is to develop a new data fusion framework rather than explore every possible model structure.

The conclusions in this thesis are also conditional on the modelling assumptions used in the simulation studies and applications. Many of the simulations are based on Gaussian random fields with Matérn covariance and relatively simple temporal dependence. In real applications, soil moisture fields may exhibit non-Gaussian features, non-stationarity and more complex temporal structure. In such settings, I would expect the main impacts to appear in parameter estimation and uncertainty calibration rather than in point predictions: Gaussian latent-field models are robust for estimating large-scale structure, but credible intervals may not be calibrated where the true process deviates strongly from the assumed form. Similarly, the XGBoost-conformal framework should retain good predictive accuracy, but may require more careful calibration if covariate shift or strong temporal dependence violate the approximate exchangeability assumptions underlying conformal prediction. A natural direction for future work in terms of the assumption is to conduct targeted sensitivity analyses, for example, by simulating from non-Gaussian or non-stationary fields, or by applying the proposed methods to more complex real datasets, to quantify how performance changes as these assumptions are relaxed.

Future work can also relax the model structure (non-separable spatio-temporal covariances, spatially varying ranges, higher-order temporal dependence), assess performance under irregular sampling, and address computational scalability for practical use. In the INLA prior setting, the PC prior is used for most simulations and real-data applications. The choice of priors can significantly impact model parameter estimation, making it one of the most critical aspects of the Bayesian Hierarchical Model (BHM). However, this is not the main focus of the current work and may be revisited in future work.

A related point concerns the use of Augmented Dickey-Fuller (ADF) tests in the exploratory

analysis. These tests are known to have low power in short time series and can behave poorly in the presence of structural breaks or strong seasonality, so their output should not be over-interpreted. In this thesis, the ADF results were used only as a rough guide to the presence of unit-root behaviour and to motivate simple differencing and low-order autoregressive components, rather than as a formal decision rule for model specification. I therefore do not expect the unusual findings from the ADF tests to materially affect the main conclusions, which are based on simulation studies and cross-validation rather than strict assumptions. However, a more systematic assessment of temporal dependence using alternative tests or model-based diagnostics would be a useful extension in future work.

Another important point is the data quality of the soil moisture measurements. The monitoring network for the in-situ sensor data is sparse and unevenly distributed throughout the study catchment. Specifically, this study's catchment only has 22 volumetric water content (VWC) sensors, and not all of them are operational at all times. Additionally, there are only 10 rainfall sensors, which are distributed along the river rather than evenly across the catchment, potentially introducing bias. An optimal deployment of the sensors could be achieved through a better sampling design. According to simulation studies, doubling the number of sensors would improve model performance, as indicated by a reduction in root mean square error (RMSE). This suggests that the current number of sensors is not enough for this study's catchment. It is noted that the recommendation to "double the number of sensors" should be interpreted as an ideal scenario rather than a strict requirement. In practice, substantially increasing the number of soil moisture sensors in a catchment may be unrealistic because of cost, maintenance and access constraints. If the network cannot be expanded to this extent, the main implication is that prediction uncertainty will remain higher in poorly instrumented parts of the catchment, especially where satellite products are also uncertain. However, there are several ways to partially compensate without a large increase in sensor numbers. One is to add a smaller number of additional sensors in a more targeted way, focusing on areas where the current uncertainty is largest. Another is to exploit additional covariates and data sources (for example, rainfall radar and land-cover) within the same fusion framework to strengthen spatial support. Finally, more flexible model structures, such as the spatially regularised XGBoost and conformal prediction framework developed in Chapter 6, can help to stabilise predictions and quantify uncertainty even when the physical sensor network is relatively sparse. Exploring these trade-offs and designing near-optimal sensor configurations is a natural direction for future work.

Furthermore, rainfall impacts soil moisture in a complex way, which indicates that a more physically based approach to modelling the rainfall covariate should be considered. For the satellite data, rainfall and air temperature measurements are not available in this study catchment, but their inclusion would be beneficial. Finally, the computational time required for the spatio-temporal data fusion model is huge, especially with a large number of time points, so it may be worthwhile to explore ways to improve efficiency.

The previous chapters point to several clear directions for future work. For the INLA-SPDE data-fusion framework, it will make the Bayesian data-fusion model more flexible. More flexible priors are tried and explored for the mesh design to balance accuracy and cost. Alternative temporal structures, including simple non-separable spatio-temporal forms, have also been explored (Cressie and Huang, 1999).

For the XGBoost with spatio-temporal locally weighted conformal prediction (stLSCP), the smoothing and weighting parameters (τ, h_s, h_t) are tuned jointly and learn the neighbour graph from data rather than fixing it in advance. Conformal calibration will be made more adaptive by letting intervals respond to covariates and by choosing bandwidths to meet a target effective sample size. Finally, this model will be evaluated over longer periods (including dry and wet extremes) with spatially and temporally blocked cross-validation, reported coverage and interval width, and tested transfer to new catchments with limited re-tuning.

The simple physics rules can also be incorporated into both learning and calibration. First, a penalty term will be included in the training loss function that discourages predictions that go against basic soil moisture behaviour. In practice, a simple diffusion residual on the grid (e.g., Richards' equation (Richards, 1931) or a reaction-diffusion equation (Tartakovsky et al., 2020)) can be computed, and its squared value can be added to the loss function with a weight. Simple bounds will be added: non-negative bounds $(0 \leq \text{VWC} \leq 1)$. The weight will be chosen by cross-validation.

Finally, performance across multiple seasons (dry and wet extremes) will be evaluated and use spatially and temporally blocked cross-validation, rather than a single day as an example, to obtain a more general measure of goodness of fit. It is also worth testing how the model transfers to other catchments with limited re-tuning to understand how well the model transfers.

Bibliography

Al-Kayssi, AW, Al-Karaghouli, AA, Hasson, AM, and Beker, SA. Influence of soil moisture content on soil temperature and heat storage under greenhouse conditions. *Journal of Agricultural Engineering Research*, 45:241–252, 1990.

Al Majou, Hassan, Bruand, Ary, and Duval, Odile. The use of in situ volumetric water content at field capacity to improve the prediction of soil water retention properties. *Canadian Journal of Soil Science*, 88(4):533–541, 2008.

Amedeo, Douglas and Golledge, Reginald G. *An Introduction to Scientific Reasoning in Geography*. John Wiley & Sons, New York, 1975.

Amemiya, Takeshi and Wu, Roland Y. The effect of aggregation on prediction in the autoregressive model. *Journal of the American Statistical Association*, 67(339):628–632, 1972.

APXML. Xgboost: Sparsity-aware. URL https://apxml.com/courses/mastering-gradient-boosting-algorithms/chapter-4-xgboost-extreme-gradient-boosting/xgboost-sparsity-aware. Accessed: 2025-06-08.

Banerjee, Sudipto, Carlin, Bradley P, and Gelfand, Alan E. *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC, 2003.

Batchu, Vishal, Nearing, Grey, and Gulshan, Varun. A machine learning data fusion model for soil moisture retrieval. *arXiv preprint arXiv:2206.09649*, 2022.

Bergmeir, Christoph and Benítez, José M. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

Berrocal, Veronica J, Gelfand, Alan E, and Holland, David M. A bivariate space-time downscaler under space and time misalignment. *The annals of applied statistics*, 4(4):1942, 2010.

Blangiardo, Marta and Cameletti, Michela. *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, 2015.

Box, George EP, Jenkins, Gwilym M, Reinsel, Gregory C, and Ljung, Greta M. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Breiman, Leo. Bagging predictors. *Machine learning*, 24:123–140, 1996.

Breiman, Leo. Random forests. *Machine learning*, 45:5–32, 2001.

Cameletti, Michela, Lindgren, Finn, Simpson, Daniel, and Rue, Håvard. Spatio-temporal modeling of particulate matter concentration through the spde approach. *AStA Advances in Statistical Analysis*, 97:109–131, 2013.

Cameletti, Michela, Gómez-Rubio, Virgilio, and Blangiardo, Marta. Bayesian modelling for spatially misaligned health and air pollution data through the inla-spde approach. *Spatial Statistics*, 31:100353, 2019.

Cawley, Gavin C and Talbot, Nicola LC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11: 2079–2107, 2010.

Chacón-Montalván, Erick A, Atkinson, Peter M, Nemeth, Christopher, Taylor, Benjamin M, and Moraga, Paula. Spatial latent gaussian modelling with change of support. *arXiv preprint arXiv:2403.08514*, 2024.

Chen, Tianqi. Xgboost: A scalable tree boosting system. *Cornell University*, 2016.

Chilès, Jean-Paul and Delfiner, Pierre. *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons, Hoboken, NJ, 2nd edition, 2012.

Cliff, Andrew David and Ord, J Keith. Spatial processes: models & applications. *(No Title)*, 1981.

Copernicus Land Monitoring Service. Soil moisture, 2024. URL https://land.copernicus.eu/en/products/soil-moisture. Accessed: 2024-11-11.

Cowles, Mary Kathryn, Yan, Jun, and Smith, Brian. Reparameterized and marginalized posterior and predictive sampling for complex bayesian geostatistical models. *Journal of Computational and Graphical Statistics*, 18(2):262–282, 2009.

Cressie, Noel and Huang, Hsin-Cheng. Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical association*, 94(448):1330–1339, 1999.

Cressie, Noel and Johannesson, Gardar. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1):209–226, 2008.

Cressie, Noel and Wikle, Christopher K. *Statistics for spatio-temporal data*. John Wiley & Sons, 2015.

Cressie, Noel A. C. *Statistics for Spatial Data*. John Wiley & Sons, 1993.

Dai, Hongbin, Huang, Guangqiu, Wang, Jingjing, and Zeng, Huibin. Var-tree model based spatio-temporal characterization and prediction of o3 concentration in china. *Ecotoxicology and environmental safety*, 257:114960, 2023.

David, Florence Nightingale and Johnson, Norman Lloyd. The probability integral transformation when parameters are estimated from the sample. *Biometrika*, 35(1/2):182–190, 1948.

De Iaco, Sandra, Myers, Donald E, and Posa, Donato. Nonseparable space-time covariance models: some parametric families. *Mathematical Geology*, 34(1):23–42, 2002.

Dickey, David A and Fuller, Wayne A. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a):427–431, 1979.

Dorigo, Wouter, Wagner, Wolfgang, Hohensinn, Roland, Hahn, Sebastian, Paulik, Christoph, Xaver, Angelika, Gruber, Alexander, Drusch, Matthias, Mecklenburg, Susanne, van Oevelen, Peter, Robock, Alan, and Jackson, Thomas. The international soil moisture network: a data hosting facility for global in situ soil moisture measurements. *Hydrology and Earth System Sciences*, 15:1675–1698, 2011.

Dormann, Carsten F, McPherson, Jana M, Araújo, Miguel B, Bivand, Roger, Bolliger, Janine, Carl, Gudrun, Davies, Richard G, Hirzel, Alexandre, Jetz, Walter, Kissling, W Daniel, et al. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, pages 609–628, 2007.

Elliott, Andrew. Home, n.d. URL https://sites.google.com/site/elliottande/home. Accessed: 2024-11-07.

Entekhabi, Dara, Rodriguez-Iturbe, Ignacio, and Castelli, Fabio. Mutual interaction of soil moisture state and atmospheric processes. *Journal of Hydrology*, 184(1–2):3–17, 1996.

European Space Agency. Sentinel-1: Radar vision for Copernicus. https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-1, 2025a. Accessed 2025-09-24.

European Space Agency. Sentinel-1. https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-1, 2025b. Accessed: 2025-01-21.

Ferreira, Marco A. R., Higdon, David M., Lee, Herbert K. H., and West, Mike. Multi-scale and hidden resolution time series models. *Bayesian Analysis*, 1(4):947–968, 2006. doi: 10.1214/06-BA131.

Forlani, Chiara, Bhatt, Samir, Cameletti, Michela, Krainski, Elias, and Blangiardo, Marta. A joint bayesian space–time model to integrate spatially misaligned air pollution data in r-inla. *Environmetrics*, 31(8):e2644, 2020.

Friedman, Jerome H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Fuentes, Montserrat and Raftery, Adrian E. Model evaluation and spatial interpolation by bayesian combination of observations with outputs from numerical models. *Biometrics*, 61(1):36–45, 2005.

Fuglstad, Geir-Arne, Simpson, Daniel, Lindgren, Finn, and Rue, Håvard. Constructing priors that penalize the complexity of gaussian random fields. *Journal of the American Statistical Association*, 114(525):445–452, 2019.

Gelfand, Alan E, Zhu, Li, and Carlin, Bradley P. On the change of support problem for spatio-temporal data. *Biostatistics*, 2(1):31–45, 2001.

Godoy, Lucas da Cunha, Prates, Marcos Oliveira, and Yan, Jun. Statistical inferences and predictions for areal data and spatial data fusion with hausdorff–gaussian processes. *arXiv preprint arXiv:2208.07900*, 2022.

Gotway, Carol A and Young, Linda J. Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648, 2002.

Gotway, Carol A and Young, Linda J. A geostatistical approach to linking geographically aggregated data from different sources. *Journal of Computational and Graphical Statistics*, 16 (1):115–135, 2007.

Gruber, Alexander, Scanlon, Tracy M., van der Schalie, René, and Wagner, Wolfgang. Evolution of the esa cci soil moisture climate data records and their role in climate research. *Remote Sensing*, 11(12):1433, 2019.

Gryparis, Alexandros, Paciorek, Christopher J, Zeka, Ariana, Schwartz, Joel, and Coull, Brent A. Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*, 10(2):258–274, 2009.

Hajibabaee, Parisa, Pourkamali-Anaraki, Farhad, and Hariri-Ardebili, Mohammad Amin. Adaptive conformal prediction intervals using data-dependent weights with application to seismic response prediction. *IEEE Access*, 12:53579–53597, 2024.

He, Shiyu and Wong, Samuel WK. Spatio-temporal data fusion for the analysis of in situ and remote sensing data using the INLA-SPDE approach. *Spatial Statistics*, 64:100863, 2024.

Hernández-Sánchez, Juan Carlos, Monsivais-Huertero, Alejandro, Judge, Jasmeet, and Jiménez-Escalona, José Carlos. Downscaling smap soil moisture retrievals over an agricultural region in central mexico using machine learning. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 7049–7052. IEEE, 2019.

Huang, Shuzhe, Zhang, Xiang, Chen, Nengcheng, Ma, Hongliang, Fu, Peng, Dong, Jianzhi, Gu, Xihui, Nam, Won-Ho, Xu, Lei, Rab, Gerhard, et al. A novel fusion method for generating surface soil moisture data with high accuracy, high spatial resolution, and high spatio-temporal continuity. *Water Resources Research*, 58(5):e2021WR030827, 2022.

James Hutton Institute. Soil data and maps, 2024. URL https://www.hutton.ac.uk/soil-data-and-maps. Accessed: 2024-11-07.

Jemeļjanova, Marta, Kmoch, Alexander, and Uuemaa, Evelyn. Adapting machine learning for environmental spatial data-a review. *Ecological Informatics*, 81:102634, 2024.

Jing, Yinghong, Li, Yao, Li, Xinghua, Lin, Liupeng, She, Xiaojun, Jiang, Menghui, and Shen, Huanfeng. An integrated learning framework for seamless high-resolution soil moisture estimation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Journel, Andre G and Huijbregts, Charles J. Mining geostatistics. 1976.

Kerr, Yann H., Waldteufel, Philippe, Wigneron, Jean-Pierre, Delwart, Steven, Cabot, Fabrice, Boutin, Jacqueline, Escorihuela, Maria Jose, Font, Jordi, Reul, Nicolas, and Gruhier, Christophe. The smos mission: New tool for monitoring key elements of the global water cycle. *Proceedings of the IEEE*, 98(5):666–687, 2010.

Keys, Robert. Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing*, 29(6):1153–1160, 1981.

Kim, Daeun, Moon, Heewon, Kim, Hyunglok, Im, Jungho, and Choi, Minha. Intercomparison of downscaling techniques for satellite soil moisture products. *Advances in Meteorology*, 2018 (1):4832423, 2018.

Krainski, Elias, Gómez-Rubio, Virgilio, Bakka, Haakon, Lenzi, Amanda, Castro-Camilo, Daniela, Simpson, Daniel, Lindgren, Finn, and Rue, Håvard. *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. Chapman and Hall/CRC, 2018.

Lam, Nina Siu-Ngan. Spatial interpolation methods: a review. *The American Cartographer*, 10 (2):129–150, 1983.

Lee, Jonghyeok, Xu, Chen, and Xie, Yao. Kernel-based optimally weighted conformal time-series prediction. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=oP7arLOWix.

Lei, Jing, Rinaldo, Alessandro, and Wasserman, Larry. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Li, Liuyang, Zhu, Qing, Lai, Xiaoming, and Liao, Kaihua. Improved downscaling of microwave-based surface soil moisture over a typical subtropical monsoon region. *Journal of Hydrology*, 627:130431, 2023.

Lin, Zhen, Trivedi, Shubhendu, and Sun, Jimeng. Conformal prediction intervals with temporal dependence. *arXiv preprint arXiv:2205.12940*, 2022.

Lindgren, Finn, Rue, Håvard, and Lindström, Johan. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

Liu, Yangxiaoyue, Xia, Xiaolin, Yao, Ling, Jing, Wenlong, Zhou, Chenghu, Huang, Wumeng, Li, Yong, and Yang, Ji. Downscaling satellite retrieved soil moisture using regression tree-based machine learning algorithms over southwest france. *Earth and Space Science*, 7(10): e2020EA001267, 2020.

Lou, Xiayin, Luo, Peng, and Meng, Liqiu. Geoconformal prediction: a model-agnostic framework of measuring the uncertainty of spatial prediction. *arXiv preprint arXiv:2412.08661*, 2024a.

Lou, Xiayin, Luo, Peng, and Meng, Liqiu. Geoconformal prediction: a model-agnostic framework of measuring the uncertainty of spatial prediction, 2024b. URL https://arxiv.org/abs/2412.08661.

Ma, Chunsheng. Spatio-temporal covariance functions generated by mixtures. *Mathematical geology*, 34(8):965–975, 2002.

Ma, Chunsheng. Spatio-temporal stationary covariance models. *Journal of Multivariate Analysis*, 86(1):97–107, 2003.

Ma, Yutiao, Hou, Peng, Zhang, Linjing, Cao, Guangzhen, Sun, Lin, Pang, Shulin, and Bai, Junjun. High-resolution quantitative retrieval of soil moisture based on multisource data fusion with random forests: A case study in the zoige region of the tibetan plateau. *Remote Sensing*, 15(6): 1531, 2023.

Madsen, Lisa, Ruppert, David, and Altman, Naomi S. Regression with spatially misaligned data. *Environmetrics: The official journal of the International Environmetrics Society*, 19(5): 453–467, 2008.

Mao, Huiying, Martin, Ryan, and Reich, Brian J. Valid model-free spatial prediction. *Journal of the American Statistical Association*, 119(546):904–914, 2024.

Mao, Taoning, Shangguan, Wei, Li, Qingliang, Li, Lu, Zhang, Ye, Huang, Feini, Li, Jianduo, Liu, Wei, and Zhang, Ruqing. A spatial downscaling method for remote sensing soil moisture based on random forest considering soil moisture memory and mass conservation. *Remote Sensing*, 14(16):3858, 2022.

McMillan, Nancy J, Holland, David M, Morara, Michele, and Feng, Jingyu. Combining numerical model output and particulate data using bayesian space–time modeling. *Environmetrics: The official journal of the International Environmetrics Society*, 21(1):48–65, 2010.

Merlin, Olivier, Rudiger, Christoph, Al Bitar, Ahmad, Richaume, Philippe, Walker, Jeffrey P, and Kerr, Yann H. Disaggregation of smos soil moisture in southeastern australia. *IEEE Transactions on Geoscience and Remote Sensing*, 50(5):1556–1571, 2012.

Meyer, Hanna and Pebesma, Edzer. Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, 13(1):2208, 2022.

Moraga, Paula, Cramb, Susanna M, Mengersen, Kerrie L, and Pagano, Marcello. A geostatistical model for combined analysis of point-level and area-level data using inla and spde. *Spatial Statistics*, 21:27–41, 2017.

Muff, Stefanie, Riebler, Andrea, Held, Leonhard, Rue, Håvard, and Saner, Philippe. Bayesian analysis of measurement error models using integrated nested laplace approximations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 64(2):231–252, 2015.

Nguyen, Hai, Cressie, Noel, and Braverman, Amy. Spatial statistical data fusion for remote sensing applications. *Journal of the American Statistical Association*, 107(499):1004–1018, 2012.

Ochsner, Tyson E., Cosh, Michael H., Cuenca, Richard H., Dorigo, Wouter A., Draper, Clara S., Hagimoto, Yuki, Kerr, Yann H., Larson, Kristine M., Njoku, Eni G., Small, Eric E., and Zreda, Marek. State of the art in large-scale soil moisture monitoring. *Soil Science Society of America Journal*, 77(6):1888–1919, 2013a.

Ochsner, Tyson E, Cosh, Michael H, Cuenca, Richard H, Dorigo, Wouter A, Draper, Clara S, Hagimoto, Yutaka, Kerr, Yann H, Larson, Kristine M, Njoku, Eni G, Small, Eric E, et al. State of the art in large-scale soil moisture monitoring. *Soil Science Society of America Journal*, 77 (6):1888–1919, 2013b.

Open-Elevation. Open-elevation: Free elevation api, 2023. URL https://open-elevation.com/. Accessed: 2023-12-02.

Paciolla, Nicola, Corbari, Chiara, Al Bitar, Ahmad, Kerr, Yann, and Mancini, Marco. Irrigation and precipitation hydrological consistency with smos, smap, esa-cci, copernicus ssm1km, and amsr-2 remotely sensed soil moisture products. *Remote Sensing*, 12(22):3737, 2020.

Peterson, Leif E. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

Pouliot, Guillaume Allaire. Spatial econometrics for misaligned data. *Journal of Econometrics*, 232(1):168–190, 2023.

Quinn, Nevil Wyndham, Newton, Chris, Boorman, David, Horswell, Michael, and West, Harry. Progress in evaluating satellite soil moisture products in Great Britain against COSMOS-UK and in-situ soil moisture measurements. Technical Report EGU2020-15831, Copernicus Meetings, March 2020.

Quiñonero-Candela, Joaquin, Sugiyama, Masashi, Schwaighofer, Anton, and Lawrence, Neil D. *Dataset shift in machine learning*. Mit Press, 2022.

Rasmussen, Carl E. and Williams, Christopher K. I. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

Reddi, Sashank, Poczos, Barnabas, and Smola, Alex. Doubly robust covariate shift correction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Richards, Lorenzo Adolph. Capillary conduction of liquids through porous mediums. *physics*, 1 (5):318–333, 1931.

Roberts, David R, Bahn, Volker, Ciuti, Simone, Boyce, Mark S, Elith, Jane, Guillera-Arroita, Gurutzeta, Hauenstein, Severin, Lahoz-Monfort, José J, Schröder, Boris, Thuiller, Wilfried, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017.

Roksvåg, Thea, Steinsland, Ingelin, and Engeland, Kolbjørn. Estimating mean annual runoff by using a geostatistical spatially varying coefficient model that incorporates process-based simulations and short records. In *EGU General Assembly Conference Abstracts*, pages EGU21–4233, 2021.

Rousseeuw, Peter J. and Croux, Christophe. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993.

Rue, Håvard, Martino, Sara, and Chopin, Nicolas. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

Sahu, Sujit K, Gelfand, Alan E, and Holland, David M. Fusing point and areal level space–time data with application to wet deposition. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 59(1):77–103, 2010.

Schmidt, Alexandra Mello and Gamerman, Dani. Temporal aggregation in dynamic linear models. *Journal of Forecasting*, 16(5):293–310, 1997.

Scott, E Marian. Framing data science, analytics and statistics around the digital earth concept. *Environmetrics*, 34(2):e2732, 2023.

Scottish Environment Protection Agency. Home, 2024. URL https://www.sepa.org.uk/. Accessed: 2024-11-07.

Scottish Environment Protection Agency. Rainfall observed tipping bucket rain gauges. https://timeseriesdoc.sepa.org.uk/api-documentation/before-you-start/what-data-are-available/, 2025. Accessed 29 September 2025.

Senanayake, Indishe P, Pathira Arachchilage, Kalani RL, Yeo, In-Young, Khaki, Mehdi, Han, Shin-Chan, and Dahlhaus, Peter G. Spatial downscaling of satellite-based soil moisture products using machine learning techniques: A review. *Remote Sensing*, 16(12):2067, 2024.

Seneviratne, Sonia I, Corti, Thierry, Davin, Edouard L, Hirschi, Martin, Jaeger, Eric B, Lehner, Irene, Orlowsky, Boris, and Teuling, Adriaan J. Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Science Reviews*, 99(3-4):125–161, 2010.

Shafer, Glenn and Vovk, Vladimir. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Shetty, Shobitha, Schneider, Philipp, Stebel, Kerstin, Hamer, Paul David, Kylling, Arve, and Berntsen, Terje Koren. Estimating surface no2 concentrations over europe using sentinel-5p tropomi observations and machine learning. *Remote Sensing of Environment*, 312:114321, 2024.

Shi, Jianbo and Malik, Jitendra. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

Simpson, Daniel, Rue, Håvard, Riebler, Andrea, Martins, Thiago G., and Sørbye, Sigrunn H. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28, 2017. doi: 10.1214/16-STS576.

Spiegelhalter, David J, Best, Nicola G, Carlin, Bradley P, and Van Der Linde, Angelika. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(4):583–639, 2002.

Stein, Michael L. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 1999.

Suen, Man Ho, Naylor, Mark, and Lindgren, Finn. Cohering disaggregation and uncertainty quantification for spatially misaligned data. *arXiv preprint arXiv:2502.10584*, 2025.

Szpiro, Adam A, Sheppard, Lianne, and Lumley, Thomas. Efficient measurement error correction with spatially misaligned data. *Biostatistics*, 12(4):610–623, 2011.

Tartakovsky, Alexandre M, Marrero, C Ortiz, Perdikaris, Paris, Tartakovsky, Guzel D, and Barajas-Solano, David. Physics-informed deep neural networks for learning parameters and constitutive relationships in subsurface flow problems. *Water Resources Research*, 56(5): e2019WR026731, 2020.

Tobler, Waldo R. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46(2):234–240, 1970.

UK Centre for Ecology and Hydrology. COSMOS-UK: The UK COsmic-ray Soil Moisture Observing System, 2024. URL https://cosmos.ceh.ac.uk. Accessed: 2024-11-11.

Villejo, Stephen Jun, Illian, Janine B, and Swallow, Ben. Data fusion in a two-stage spatio-temporal model using the inla-spde approach. *Spatial Statistics*, 54:100744, 2023.

Villejo, Stephen Jun, Martino, Sara, Lindgren, Finn, and Illian, Janine B. A data fusion model for meteorological data using the inla-spde method. *Journal of the Royal Statistical Society Series C: Applied Statistics*, page qlaf012, 2025.

Vovk, Vladimir, Gammerman, Alexander, and Shafer, Glenn. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Wackernagel, Hans. *Multivariate Geostatistics: An Introduction with Applications*. Springer-Verlag, 3 edition, 2003.

Wang, Craig and Furrer, Reinhard. Combining heterogeneous spatial datasets with process-based spatial fusion models: A unifying framework. *arXiv preprint arXiv:1906.00364*, 2019.

Wang, Zidong, Wu, Xianhua, and Wu, You. A spatiotemporal xgboost model for pm2. 5 concentration prediction and its application in shanghai. *Heliyon*, 9(12), 2023.

Wei, Zushuai, Meng, Yizhuo, Zhang, Wen, Peng, Jian, and Meng, Lingkui. Downscaling smap soil moisture estimation with gradient boosting decision tree regression over the tibetan plateau. *Remote Sensing of Environment*, 225:30–44, 2019.

Weisberg, Sanford. Yeo-johnson power transformations. *Department of Applied Statistics, University of Minnesota. Retrieved June*, 1:2003, 2001.

Western, Andrew W. and Blöschl, Günter. On the spatial scaling of soil moisture. *Journal of Hydrology*, 217(1–2):203–224, 1999.

Wikipedia contributors. Integer overflow, 2025. URL https://en.wikipedia.org/w/index.php?title=Integer_overflow. Accessed: 2025-01-12.

Wikle, Christopher K and Berliner, L Mark. Combining information across spatial scales. *Technometrics*, 47(1):80–91, 2005.

Wilkie, CJ, Miller, CA, Scott, EM, and O'Donnell, RA. Nonparametric statistical downscaling for the fusion of data of different spatiotemporal support. *Environmetrics*, 30(3):e2549, 2019.

Wilson, Katie and Wakefield, Jon. Pointless spatial modeling. *Biostatistics*, 21(2):e17–e32, 2020.

Wong, Pei-Yi, Lee, Hsiao-Yun, Chen, Yu-Cheng, Zeng, Yu-Ting, Chern, Yinq-Rong, Chen, Nai-Tzu, Lung, Shih-Chun Candice, Su, Huey-Jen, and Wu, Chih-Da. Using a land use regression model with machine learning to estimate ground level pm2. 5. *Environmental Pollution*, 277: 116846, 2021.

Yang, Chenconghai, Yang, Lin, Zhang, Lei, and Zhou, Chenghu. Soil organic matter mapping using inla-spde with remote sensing based soil moisture indices and fourier transforms decomposed variables. *Geoderma*, 437:116571, 2023.

Zapata-Marin, Sara, Schmidt, Alexandra M, Weichenthal, Scott, and Lavigne, Eric. Modeling temporally misaligned data across space: The case of total pollen concentration in Toronto. *Environmetrics*, page e2820, 2023.

Zhong, Ruiman and Moraga, Paula. Bayesian hierarchical models for the combination of spatially misaligned data: a comparison of melding and downscaler approaches using inla and spde. *Journal of Agricultural, Biological and Environmental Statistics*, pages 1–20, 2023.

Zhong, Ruiman, Ribeiro Amaral, André Victor, and Moraga, Paula. Spatial data fusion adjusting for preferential sampling using integrated nested laplace approximation and stochastic partial differential equation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 188(1):140–157, 2025.

Zhou, Shijie and Bradley, Jonathan R. Bayesian hierarchical modeling for bivariate multiscale spatial data with application to blood test monitoring. *Spatial and Spatio-temporal Epidemiology*, 50:100661, 2024.

Zhou, Wenbin, Zhu, Shixiang, Qiu, Feng, and Wu, Xuan. Hierarchical spatio-temporal uncertainty quantification for distributed energy adoption. *arXiv preprint arXiv:2411.12193*, 2024.