Data-Driven Spatiotemporal Modeling in the Geosciences using Machine Learning

Master Thesis of

Nicholas Chandler

Matriculation Number: 106468

Berlin University of Applied Sciences and Technology (BHT)
Fachbereich VI - Informatik und Medien
Cognitive Algorithms Lab (Calgo Lab)

Reviewer:	Prof. Dr. Felix Bießmann
Second reviewer:	Prof. Dr. Florian Hruby

01.04.2026 – 01.09.2026

Berlin University of Applied Sciences and Technology (BHT)
Fachbereich VI - Informatik und Medien
Luxemburger Straße 10
13353 Berlin

Kurzfassung

Die raumzeitliche Modellierung findet in vielen Teilbereichen der Geowissenschaften Anwendung. In dieser Arbeit konzentrieren wir uns auf die Anwendung der raumzeitlichen Modellierung auf den Grundwasserspiegel in Brandenburg. Die Modellierung im räumlichen, zeitlichen und raumzeitlichen Bereich erfolgt mithilfe von maschinellen Lernverfahren und Gaußprozessen auf Basis eines historischen Datensatzes von Grundwasserspiegeln sowie geologischen, umweltbezogenen und fernerkundlichen Kovariaten. Die Ergebnisse umfassen einen Vergleich modernster neuronaler Prognosemodelle mit auf Zeitreihen Foundation Models, eine Bewertung der räumlichen Interpolationsfähigkeit von maschinellen Lernverfahren und Gaußprozessen sowie die Entwicklung mehrerer neuartiger raumzeitlicher Modellierungsansätze. Zusätzlich wird die Modellleistung in Bezug auf die geologische und geografische Region mittels Hauptkomponentenanalyse statischer Kovariaten und geografischer Karten untersucht, um Merkmale zu identifizieren, die mit hoher bzw. niedriger Modellleistung korrelieren. Die Arbeit schließt mit einem Vergleich des besten raumzeitlichen Modells mit Basismodellen, um seine Leistungsfähigkeit im Hinblick auf zwei wichtige Vergleichspunkte besser zu verstehen.

Abstract

The work of spatiotemporal modeling is useful in many subdomains of the geosciences. In this thesis, we draw our attention to the application of spatiotemporal modeling to the groundwater level in Brandenburg, Germany. Modeling in the spatial, temporal, and spatiotemporal domain are done using machine learning based models and Gaussian processes given a historical dataset of groundwater levels, geologic, environmental, and remote sensing covariates. The results consist of a comparison between state of the art neural forecast models and time series foundation models, an evaluation of the spatial interpolation ability of machine learning and Gaussian process methods, and the development of several novel spatiotemporal modeling approaches. Additionally, the investigation of model performances with respect to the geologic and geographic region using principal components analysis of static covariates and geographic maps is given to identify which features may correlate with high and low model performances. The work closes with a comparison of the best spatiotemporal model with baselines to better understand its performance relative to two important points of comparison.

Acknowledgements

This thesis was made possible by many people and the supporting systems which allowed it to come to fruition. Namely, I want to thank my advisors and reviewers for allowing me to meet with them and for their guidance and assistance throughout. I want to thank the lab where I conducted the majority of the work and the BHT for providing an environment condusive to success. I want to thank the collaborators at BGR and Lawrence Berkeley National Lab. Finally, I want to thank my family for their continued support through my academic journey.

Contents

List of Figures

List of Tables

1. Introduction

This work concerns the development and evaluation of machine learning methods used for spatiotemporal modeling in the geosciences, with a particular focus on groundwater levels. To do this over 70 years of sensor data from monitoring wells throughout Brandenburg, Germany are considered. In addition, a focus is placed on the use of machine learning based, neural models for forecasting and Gaussian process models for spatial interpolation. The motivation, goals, and structure of the remainder of the thesis are outlined in the following sections.

1.1. Motivation

Groundwater is an important resource for many reasons. For example, approximately 2.5% of the water on Earth is freshwater. Of that, approximately 30% is groundwater with the largest portion of this freshwater in the form of glaciers and ice caps [39]. Due to this, groundwater is withdrawn in large quantities from reservoirs around the world. In 2015 in the USA, for example, on average about 8.46 billion gallons per day of groundwater were withdrawn. The uses of this water were numerous including irrigation, public supply, mining, and livestock to name a few [29]. Another example of the importance of groundwater is its use as a source of drinking water. Approximately 80% and 50% of the drinking water is groundwater in the EU and USA, respectively [45].

Due to this reliance on groundwater, as depicted in the above examples, it is important to have effective methods for the modeling of groundwater levels. The modeling can be decomposed into temporal forecasting and spatial interpolation. That is, forecasting future groundwater levels at a site given past measurements and determining the groundwater levels at locations where there are no monitoring wells. Forecasting of groundwater levels has been done using a variety of methods with more recent work utilizing machine learning methods and neural networks as in Kunz et al. [27]. Additionally, spatial interpolation of groundwater levels has been done using methods such as kriging as in Kumar et al. [26].

There have been few attempts to combine both forecasting and interpolation to do spatiotemporal modeling of groundwater in the geosciences using machine learning. The main motivation for this work is to investigate the application of foundation models and spatial models to the problems of groundwater level forecasting and spatial interpolation. The result of having an effective spatiotemporal model of groundwater levels is that forecasts of groundwater levels could be made at sites where there are no monitoring wells. This would benefit society in the sense that it would facilitate better management of groundwater resources and allow for better planning of water usage. Additionally, the results of this work benefit the machine learning community by providing another instance of spatiotemporal modeling and evaluating the procedures used therein on a novel dataset.

1.2. Goals of the Research

The goals of this research are manyfold, but primarily, this work aims to develop and evaluate a spatiotemporal model that is capable of forecasting groundwater levels at sites where there are no monitoring wells. In order to achieve this, the following research questions will be investigated:

Core Questions

1. Can foundation models (zero-shot or finetuned) be used for groundwater level forecasting?
2. Can decoupled spatiotemporal models with forecast-interpolate steps effectively forecast the groundwater level at unmonitored sites?
3. Do coupled spatiotemporal models with joint training of the spatial and temporal aspects outperform decoupled models in groundwater level forecasting?
4. Do geologic covariates improve the predictive performance?
5. Which geologic covariates correlate with higher/lower performance?

1.3. Thesis Structure

In the remainder of the thesis, the following structure is used. We begin with related work examining basic general hydrology, methods used in the literature in support of our goals, methods that we decided not to consider, the development of specific methods in the machine learning community, and examples of spatiotemporal modeling in the geosciences. We then move to the methodology where the data including sources, preprocessing, exploratory analysis and splits are examined. Next, modeling is discussed beginning with solely temporal models, then solely spatial models, and finally spatiotemporal models. The results section provides experimental results for each of the modeling scenarios. Thereafter a discussion section explores the results of the previous section. The thesis comes to a close with the conclusion where the limitations, core questions, and future work are discussed. In more detail, the thesis structure is outlined in the following subsections.

1.3.1. Related Work

The related work section covers some of the fundamental points on hydrology, forecasting, spatial modeling, and spatiotemporal modeling. Additionally, it covers some applications concerning the problem domain of groundwater modeling and other environmental science applications.

1.3.2. Data

The data section covers the dataset we use, mentioning preprocessing and other characteristics.

EDA There is exploratory data analysis where statistics are computed and the time series are visualized. The autocorrelation function of the target with respect to time is visualized along with the cross correlation of the target variable with respect to distances. Finally, static and dynamic covariates are discussed in a table.

PCA Principal Components Analysis (PCA) is done to better understand the static covariates of the dataset and to mitigate computational issues with Gaussian process kernels that arise with multicollinear features. A description, scree plot, and two relevant biplots as well as interpretations are provided.

Data Splits To ensure we appropriately measure the generalization performance of the models tested, a data splitting scheme is needed and therefore developed. The method to do this while respecting time and space is given along with the resulting counts of the sub-datasets. Finally, a geographic map visualization of the spatial distribution of wells in each of the spatial splits is provided.

1.3.3. Methods

The following sections detail the target problems and the methodology used to investigate the problems.

Temporal The temporal section is concerned with forecasting the groundwater level at sites which have historical groundwater levels. The models used are: the Naive Baseline, N-HiTS, Chronos2 Zero-Shot, and Chronos2 Finetuned. The forecasts are distributional forecasts in the cases of the neural forecasters.

Spatial The spatial section seeks to use the historical groundwater level at known sites to predict the mean groundwater level at unknown sites. The models used are: K-NN, TabPFN, Residual Neural Networks, and Gaussian Processes. Where possible, the distributional prediction performance is evaluated.

Spatiotemporal The spatiotemporal portion explores the use of historical groundwater levels at known sites to do distributional forecasting of the groundwater level at sites with unknown historical groundwater levels. To do this, the portion is split methodologically into decoupled and coupled spatiotemporal sections. The decoupled section focuses primarily on the application of the Gaussian process models to the forecasts of each of the mentioned temporal forecasters, in effect, interpolating the forecasts. The coupled section uses the Kronecker Gaussian process where the structure of the data is exploited to yield computational feasibility of a direct exact Gaussian process model, whereas without, this would not be possible. The coupled section also uses attention with a quantile output module to do forecasting by attending over learned neural representations of the spatial features at sites with known and unknown historical groundwater levels as well as the forecasted groundwater levels at sites with known histories. The final model tried passes the attended representation into a variational Gaussian process.

Evaluation Procedure This section briefly details the data splits used in each of the aforementioned tasks and documents the metrics used for evaluation.

1.3.4. Results

The results section examines the hyperparameters and model performances. Additionally, it explores the contrast between the Gaussian processes on N-HiTS with and without geologic principal component information.

Models and Hyperparameters For each of the models mentioned in the methods section, the optimal hyperparameters are recorded and the performance on the metrics are given.

Best Spatiotemporal Models & Relative Comparisons The analysis of the best spatiotemporal model is done to examine how the performance varies in geographic and principal component space. Additionally, comparison to a baseline expected to perform worse and comparison to a baseline expected to perform better is done to see how the model behaves relative to two scenarios of interest.

1.3.5. Discussion

The discussion section covers interpretations of the results between the previous sections.

1.3.6. Conclusion

The conclusion section closes the thesis by addressing the research questions, exploring limitations, and considering future work.

Research Goals This section revisits the core questions outlined previously with the knowledge developed during the thesis.

Limitations and Future Work The limitations and future work section identifies shortcomings of the thesis, potential areas for improvement, unexpected results, and the directions that could be worth exploring next, given the body of work just completed.

2. Related Work

In this section, we discuss fundamentals in hydrology, work in time series forecasting, spatial interpolation methods, and applications in the geosciences. Special attention is drawn to foundation models for time series forecasting and Gaussian processes for spatial modeling.

2.1. Hydrology

Hydrology is a subfield of geology and more generally environmental sciences concerned with the study of water on Earth. Water on Earth is governed by the general water balance equation [8]:

$$P = Q + ET + \Delta S \tag{2.1}$$

where:

- P is precipitation
- Q is streamflow
- ET is evapotranspiration
- ΔS is the change in storage

One can separate the field of hydrology into surface and subsurface hydrology, the latter is our primary concern. Groundwater is the primary interest and is defined as water occuring below the surface in porous rock [43]. In the above equation, groundwater is primarily associated with the ΔS term and is, as stated, affected by the other components.

The movement of groundwater is governed by Darcy's Law [14]:

$$q = -K\nabla h \tag{2.2}$$

where:

- q is the specific discharge
- K is the hydraulic conductivity
- ∇h is the hydraulic gradient, representing the spatial change in total hydraulic head h.

Groundwater takes time to flow into and *recharge* aquifers, which depends on the depth of the aquifer, the terrain, and the material between the surface and the aquifer. This recharge time is accounted for by utilizing a 52-week context window that is investigated in more detail in Section 3.1. Additionally, Darcy's law says the rate of flow depends on hydraulic conductivity, K which is not necessarily smooth throughout space [14]. To account for this, and since it may be informed by land cover or geologic covariates, they are included and investigated in the modeling.

Additionally, the notion of confined and unconfined aquifers is a relevant hydrological concept for this work. Essentially an unconfined aquifer is one in which the highest depth is the water table and where no aquitard sits above the aquifer. Confined aquifers are those seated below a confining bed (aquitard) which can then generate pressure such that when a well is drilled, the surface of the water is forced upwards, sometimes above the water table. A well of this type would be called an *artesian well.* More detail is given in Woessner and Poeter [43].

In the following work, we utilize primarily machine learning approaches that model the behavior of groundwater spatiotemporally given data deemed relevant to the aforementioned equations by hydrologists associated with the German Federal Institute for Geosciences and Natural Resources. These data are expanded upon and explored further in Section 3.1.

2.2. Forecasting

Time series forecasting is defined as the task of predicting future values of a time-varying signal given historical values [20]. Time series forecasting in the data-driven sense is a domain that has had interest since the 1950s [9]. It has seen development with statistical models such as Auto Regression, Moving Averages, ARIMA, and SARIMAX among others [20]. These models are primarily linear and tend to be limited in representative power when compared to modern neural architectures [18]. Neural architectures have developed from Long Short Term Memory (LSTM) units [15], to Gated Recurrent Units (GRUs) [7], and most recently Transformer architectures [42]. LSTMs suffer from autoregressive error accumulation and transformers suffer from high computational cost ($O(T^2)$ memory for a sequence length of T) and need for lots of data which motivated the development for more efficient, yet still expressive architectures. With these building blocks, modern neural forecasting architectures such at Temporal Fusion Transformers (TFT) [28] and N-HiTS [6] were developed. In this work, N-HiTS is utilized due to the state of the art predictive performance on the task of groundwater level forecasting in Germany Kunz et al. [27].

In a separate vein, foundation models have seen great success in the domains of computer vision and natural language processing [20]. In essence, foundation models are large neural architectures that are given many examples of different datasets for pre-training with the intent of having a representation of a variety of tasks within one model [44]. Additionally, foundation models can be finetuned for specific tasks to enhance their downstream performance with the context of the diverse pre-training data collection. This paradigm has seen application in time series forecasting as well. An early time series foundation model was created by training N-BEATS on many time series to enable zero shot capabilities [20]. Since then there has been interest in the development of foundation models for both zero shot and finetuned time series forecasting.

Notable examples of time series foundation models include Lag-Llama [36], TabPFN-TS [17], and Chronos2 [2]. TabPFN-TS is a modified TabPFN (Section 2.3) architecture which ensures the non-exchangability of time series data is accounted for. TabPFN-TS is not utilized here due to the closed-source code, less permissive license, and the potential inefficiency with a large number of features. In this work, Chronos2, a modified T5 architecture, is utilized due to its ability to do multivariate probabalistic forecasting, cross learning, accept past-only covariates, known future covariates, and categorical covariates. Additionally, Chronos2 has a scaling of memory that is linear with the number of variates and has support for LoRA, which allows for finetuning of a smaller number of parameters by utilizing low rank adaptations of the weight matrices instead of full-parameter updates [19]. This allows for the exploration of the applicability of foundation models in the zero shot and finetuned settings and the competitiveness on this task against the currently used neural forecasters.

2.3. Spatial Interpolation

Spatial interpolation for our purposes is defined as the task of predicting spatially continuous values at locations that are not sampled [30]. The development of spatial interpolation methods arises partly from the applicability to problems in the geosciences. This is demonstrated through the development of kriging for use in mining geology to determine the distribution of gold in South Africa [25]. Many of the folowing spatial interpolation methods make use of Tobler's first law of geography: "Everything is related to everything else, but near things are more related than distant things" [41]. That is, they utilize the distance from observed locations in making predictions at unobserved locations.

The simplest approach to spatial interpolation is the K Nearest Neighbors algorithm which uses an average of a fixed number, K, of values nearby as the prediction at unsampled locations [30]. Another approach is the Inverse Distance Weighting (IDW) algorithm which predicts unsampled locations to be a weighted average of observed values with the weights as the reciprocal of the distance to the prediction [30]. A more complex approach is the use of Gaussian Processes (GPs) where a distribution of functions is calibrated with the observed points, thus defining a set of possible functions that can fit the underlying process [35]. In the geostatistical community, kriging, as mentioned above, is what Rassmussen and Williams [35] define as Gaussian Processes. For simplicity, we utilize Gaussian Processes in this work.

Machine learning methods can also be used in spatial contexts. A notable development is TabPFN, a Prior-Data Fitted Network that approximates Bayesian inference for tabular datasets in a single forward pass using a Transformer architecture [12]. TabPFN and other neural architectures can ingest spatial coordinates alongside environmental covariates.

Additionally, a residual network (ResNet) architecture is used to do spatial interpolation [44]. However, standard deep learning models treat spatial coordinates as exchangeable features, ignoring explicit spatial graphs or localized spatial variations [33]. In an attempt to remedy this, feature expansion using randomly generated values and sinusoids as in Rahimi and Recht [34] is done.

2.4. Spatiotemporal Modeling & Applications in the Geosciences

Spatiotemporal modeling can be approached in several ways, including forecasting then interpolating as we examine in subsequent chapters, interpolating, then forecasting as in Nag et al. [31], and jointly forecasting and interpolating which we also examine.

Methods that do interpolation, then forecasting are those such as Nag et al. [31] where basis functions are used to encode spatiotemporal information about given locations before being reduced to quantile predictions by a neural network. Thereafter, forecasts are made at these interpolated points using an LSTM-architecture. This has been applied to modeling PM-2.5 in the United States [31].

Forecasting then interpolation has been applied as well as in Tapoglou et al. [40]. The authors use a neural network to model the temporal portion and then kriging to model the spatial variation, combining the approaches using fuzzy logic methods. The problem domain is groundwater level simulation.

Joint models are widespread and of many styles. A direction that parallels the approaches we cover are graph neural networks for groundwater level forecasting. Specifically, Bai et al. [3] use Graph Wave Net to do forecasting at various sites in British Columbia. While this approach is becoming more common in the literature, integration of new sites to train and inference on is more difficult than when Gaussian processes are used as the spatial interpolators. Additionally, inference on a raster spatial grid is more challenging. Finally, an aim of this thesis is to extend part of the work of [27] by conducting spatial interpolation on the forecasted values, that is doing data-driven spatiotemporal modeling in the geosciences using machine learning.

3. Methodology

This section details the datasets used and the technical aspects of the modeling applied to the data. In the data section, exploratory analysis, preprocessing, target definition, and definition of splits are done. In the modeling section, technical details about the models used for each of the tasks as well as details on evaluation methods are given.

3.1. Data

The data utilized are created by groundwater measuring wells distributed throughout Brandenburg, Germany with weekly measurement histories dating to 1951 sourced from the “Bundesanstalt für Geowissenschaften und Rohstoffe” (BGR) (German) or “Federal Institute for Geosciences and Natural Resources” (English). Specifically, these wells measure the groundwater level in meters above NormalenHöhenNull (NHN) which denotes the meters above the standard elevation zero. The variable of interest for modeling is the groundwater level and at each site, there are historical data which accompany this measurement. Additionally, there are static and dynamic covariates at each site such as the distance to streams and geomorphological uniformity as well as the time-varying temperature and precipitation at that location.

In the subsequent sections, analysis focuses on the German state of Brandenburg with a history from 1951 to 2024. Computation of descriptive statistics, creation of visualizations, and description of data splitting are given. It is noteworthy that all map visualizations use the Universal Transverse Mercator (UTM) Zone 33N projection with the European Terrestrial Reference System 1989 (ETRS89) datum (EPSG:25833). In the remainder of this section, we examine the temporal histories of the time series in the dataset, do preprocessing, compute descriptive statistics, do principal components analysis, split the data, and visualize the spatial distribution of the data using maps.

3.1.1. Preprocessing & Exploratory Analysis

While the data has coverage from 1951 to 2024, not all wells have measurements through this time period. In order to better understand the coverage of each time series, the history of each one is examined in Figure 3.1. Each row in the heatmap represents one well and each column represents a point in time. The black area represents points in time where the given well has a recorded groundwater level. As shown in the figure, all wells in the dataset are continuously operational past November 11th, 2007. At this point until the end of the data, there are approximately 16 years of weekly observations. Since the spatial data appears to be relatively sparse and there is a sufficient history to create temporal training, validation, and test splits, the data is truncated at 2008-01-01, discarding all measurements before then. This subset is the one considered in the remainder of the work.

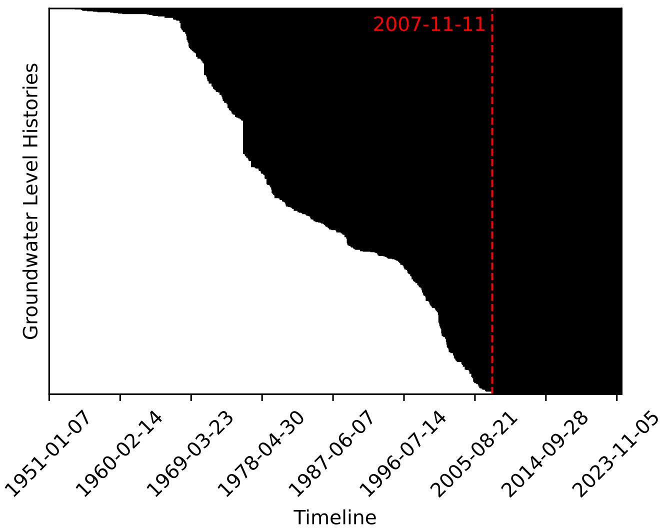

Figure 3.1.: Operational Times of Wells in Brandenburg. Black denotes wells and dates with a groundwater level recording, white denotes no recording.

As a further preprocessing step, the wells marked "ungespannt", unconfined (as opposed to confined) are utilized in the subsequent analysis and modeling. In confined aquifers, groundwater levels may reflect hydraulic pressure rather than the local water table, making them less physically comparable to unconfined levels [43]. These unconfined wells will serve as our locations for the spatial portion of the spatiotemporal modeling. Finally, there were no NAs in the groundwater level or the selected covariates after the time series beginning and additional preprocessing was done before acquisition as detailed in Kunz et al. [27].

To understand the statistical distribution of the dataset at hand, descriptive statistics are given in Table 3.1. Notably, the overall standard deviation is smaller than the mean which indicates a small spread across the spatiotemporal space.

Table 3.1.: Descriptive Statistics of Unconfined Wells

Statistic	**Computed Value**
Total number of wells	1,040
Number of unconfined wells	721
Number of wells eliminated	319
Unconfined Aquifer Levels (n=721)	
Mean groundwater level	38.627 m NHN
Median groundwater level	36.840 m NHN
Standard deviation	22.711 m NHN

Seasonality and trend are examined through the use of a heatmap visualization and an autocorrelation function plot. In Figure 3.2, one sees the groundwater levels for each well over the entire time period, standardized with respect to the mean groundwater level at each well. Each row contains the entire historical groundwater level for a given measurement well, and, correspondingly, each column is one timestep. There appears to be strong correlation between the different sites indicated by the similar values between sites at each timestep. Additionally, there appears to be some periodicity with higher-than-average values in January and lower-than-average values in June.

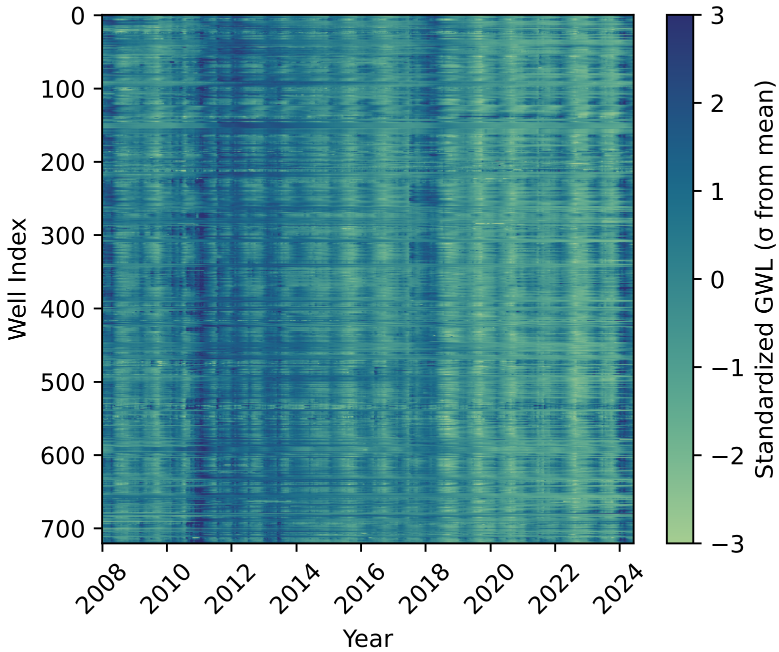

Figure 3.2.: Site Statistics Standardized Groundwater Levels in Brandenburg

The final temporal analysis is the autocorrelation function plot in Figure 3.3 This plot shows the autocorrelation function averaged across wells with lags from 0 weeks to 104 weeks. A peak occurs at lag 49 with correlation 0.538 indicating an approximately yearly seasonality. This figure indicates that there is a strong autocorrelation and so the forecasting aspect should obtain useful information from prior values of the target series. Additionally, the peak indicates that a context of 52 weeks could suffice to ensure we capture the signal contained in the given peak.

The final figure in the exploratory data analysis is Figure 3.4 which depicts the pairwise cross-correlation at specific distances. The x-axis shows the distance between wells, the y-axis shows the Pearson correlation coefficient between the timeseries at these wells, and the hexagons on the plot surface depict the binned density at each point. The largest density of well pairs appears to be above a correlation of 0.5 and near 100 kilometers of distance between the wells. The calculated average distance between two wells in the dataset is 89.342 kilometers and the mean correlation of measured groundwater levels is 0.536.

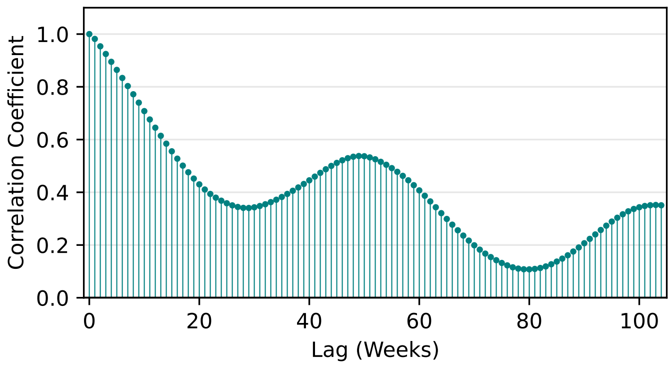

Figure 3.3.: Autocorrelation Function of Groundwater Levels in Brandenburg. Peak at approximately lag 49 with a value of 0.538.

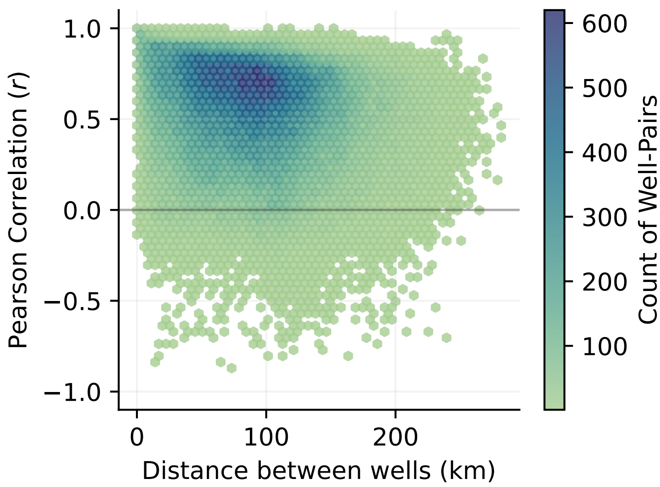

Figure 3.4.: Cross-Correlation by Distance of Wells in Brandenburg.

Covariates The following covariates are the static and dynamic covariates used in modeling and are detailed in Table 3.2. The static covariates are used in the PCA and subsequent models in the following sections though some such as the CORINE [10] and HÜK250 features are left out of the PCA. The dynamic covariates are assumed to be forecasts and are used only by the temporal models.

Table 3.2.: Overview of Metadata, Target Variables, and Model Features Grouped by Category

Parameter	Explanation	Feature Type
Metadata & Target Variable		
date	Date	-
id	Id of the monitoring well	-

Table 3.2.: Overview of Metadata, Target Variables, and Model Features Grouped by Category

Parameter	**Explanation**	**Feature Type**
gwl	Groundwater level (m NHN)	Dynamic
Topographic & Hydrologic Features		
Topographic Wetness Index (1km)	Potential soil moisture / water accumulation based on terrain	Static
Divide to Stream Distance (1km)	Divide to stream distance (Nölscher et al. [32])	Static
Lateral Position to Stream (1km)	Lateral position (Nölscher et al. [32])	Static
Stream Distance (1km)	Stream distance (Nölscher et al. [32])	Static
Groundwater Recharge Rate	Mean annual groundwater recharge rates 1961-1990 ($\mathrm{mm\,a^{-1}}$)	Static
Landscape Diversity Metrics		
Landscape Entropy (10km radius)	Mean entropy (Amatulli et al. [1])	Static
Landscape Shannon Diversity (10km radius)	Mean Shannon diversity index (Amatulli et al. [1])	Static
Landscape Unique Cover Classes (10km radius)	Mean number of unique land cover classes (Amatulli et al. [1])	Static
CORINE Land Cover Fractions		
clc_111	Continuous urban fabric	Static
clc_112	Discontinuous urban fabric	Static
clc_121	Industrial or commercial units	Static

Table 3.2.: Overview of Metadata, Target Variables, and Model Features Grouped by Category

Parameter	Explanation	Feature Type
clc_131	Mineral extraction sites	Static
clc_141	Green urban areas	Static
clc_211	Continuous urban fabric	Static
clc_231	Pastures	Static
clc_242	Complex cultivation patterns	Static
clc_311	Broad-leaved forest	Static
clc_312	Coniferous forest	Static
clc_313	Mixed forest	Static
clc_321	Natural grasslands	Static
clc_322	Moors and heathland	Static
clc_411	Inland marshes	Static
clc_511	Water courses	Static
clc_512	Water bodies	Static
Hydrogeological Permeability (HÜK250)		
huek_kf_9	Medium	Static
huek_kf_10	Low	Static
huek_kf_11	Highly variable	Static
Meteorological & Temporal Features		
day_sin	Sine of day of year (for seasonal cycle encoding)	Dynamic
day_cos	Cosine of day of year (for seasonal cycle encoding)	Dynamic
tas_5km	Mean temperature (5 km raster resolution)	Dynamic
hurs_5km	Relative humidity (5 km raster resolution)	Dynamic
pr_5km	Precipitation (5 km raster resolution)	Dynamic

Principal Components Analysis Principal Components Analysis (PCA) is an unsupervised learning technique that allows for dimensionality reduction and general data analysis. Intuitively, it represents a change of coordinates of the data matrix, $X \in \mathbb{R}^{n \times d}$ such that the principal component axes follow the directions of largest variance in the feature space. Formally, The Singular Value Decomposition (SVD) can be used to construct the principal components in Principal Component Analysis (PCA) as follows: Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a data matrix where each of the n rows represents a (feature-wise) mean-centered observation of dimension d. The SVD of $\mathbf{X}$ is defined as:

$$\mathbf{X} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^* \tag{3.1}$$

From this decomposition, the product $\mathbf{V}\boldsymbol{\Sigma}$ contains the scores of the rows of $\mathbf{X}$ (i.e., the transformed coordinates of each observation), and the columns of $\mathbf{U}$ form the matrix of the principal component loading vectors [13]. The principal component loading vectors and scores can be interpreted using biplots where the loading vectors are the directions of the features in principal component space and the scores are the postitions of observations in this space.

To maximize spatial interpretability and avoid the mathematical artifacts associated with high-dimensional compositional data, highly sparse land-cover (CORINE) and hydrogeological (HÜK250) fraction components were excluded from the spatial decomposition. Standardizing the feature space to 14 continuous morphometric and hydrographic variables yielded a highly parsimonious PCA space, where the first two principal components successfully retain approximately 50.5% of the total variance. This reduction eliminates multicollinearity of these covariates in further models due to the orthogonality of the principal components and reduces noise in the feature space. To characterize this dimensionality reduction, we examine the PCA variance plots in Figure 3.5 where 8 PCs explain 97.4% of the variance in the data. We present two biplots and interpret the principal component axes and areas in the space with respect to our dataset and features in Figure 3.6 and Figure 3.7. Namely principal components 1 and 2 are utilized in the first biplot due to the large proportion of variance explained. Principal components 1 and 8 are utilized in the second biplot due to the importance in modeling further on.

For each biplot we interpret the directions of loading vectors [13]. The Topographic Wetness Indicator (TWI) variables loading vectors follow the negative PC1 direction, indicating that larger values of PC1 have smaller values of TWI and therefore have lower topographic wetness. The hydrologic position loading vectors follow PC2, indicating that larger values of PC2 indicate farther distances from streams. Finally, the closest covariate loading vectors to the PC8 axis are the elevation in the positive direction and the groundwater recharge rate in the negative direction. This indicates larger values of PC8 correspond with higher elevation areas with a lower groundwater recharge rate and smaller values with the opposite. Finally, it is noteworthy that the PCA was fit on the Train-Train data and only the Train-Train points are visualized in this section. More details to data splitting are in Subsection 3.1.2

GeoTessera Embeddings GeoTessera embeddings compress a full year of multi-spectral and radar satellite imagery into dense, 128-dimensional vectors, providing a comprehensive

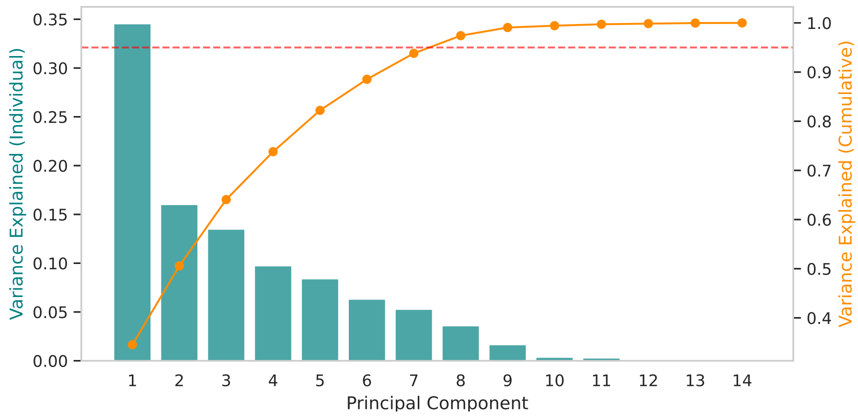

Figure 3.5.: Static Covariate PCA Variance Plot. The first two principal components contain 50.5% of the variance, 8 principal components contain 97.4%.

representation of surface reflectance and structure for coordinate points on Earth [11]. We utilize the embeddings in specific Gaussian process models where land cover components are not included as an additional spatial covariate that is static in time. In order to address the issue of incomplete temporal coverage of the GeoTessera embeddings, we obtain the embeddings at each location in our dataset for the year 2017 which falls within the temporal training set and near the center of the considered time period. These values are treated as static through time in the entire dataset. Additionally, the 128-dimensional representation has noise and in order to reduce this and the dimensionality, we apply PCA on the Train-Train set. After applying PCA, we find that 20 principal components explain 79% of the variance of the full dimensionality and use this representation in several models. Finally, the cumulative proportion of variance explained plot and biplot are included in the Appendix B.1.

Summary In this subsection, we discussed the characteristics of the Brandenburg dataset and did some preprocessing for further analysis. The dataset was temporally subsetted to include only times when there were measurements for all series, yielding history from 2008-01-01 to 2024-06-03. Thereafter, analysis on the spatial distribution was conducted and discussion about as well as subsequent removal of monitoring wells for confined aquifers was undertaken. A general analysis of the target, groundwater levels, through time and across each site was carried out with the use of visualizations. The autocorrelation function plot was used to determine the usefulness of previous groundwater levels in forecasting future values and also that a context of 52 timesteps would capture the signal included in the peak in the plot. The cross correlation of the groundwater levels in pairs of wells was examined with respect to the distance between the wells and the overall Pearson's correlation coefficient was computed to quantify the spatiotemporal relationship present in the data. The covariates were discussed with more details in the appendix. PCA on static covariates from the BGR was conducted and the biplots examined. GeoTessera embeddings were introduced and explanation of the application in our modeling was

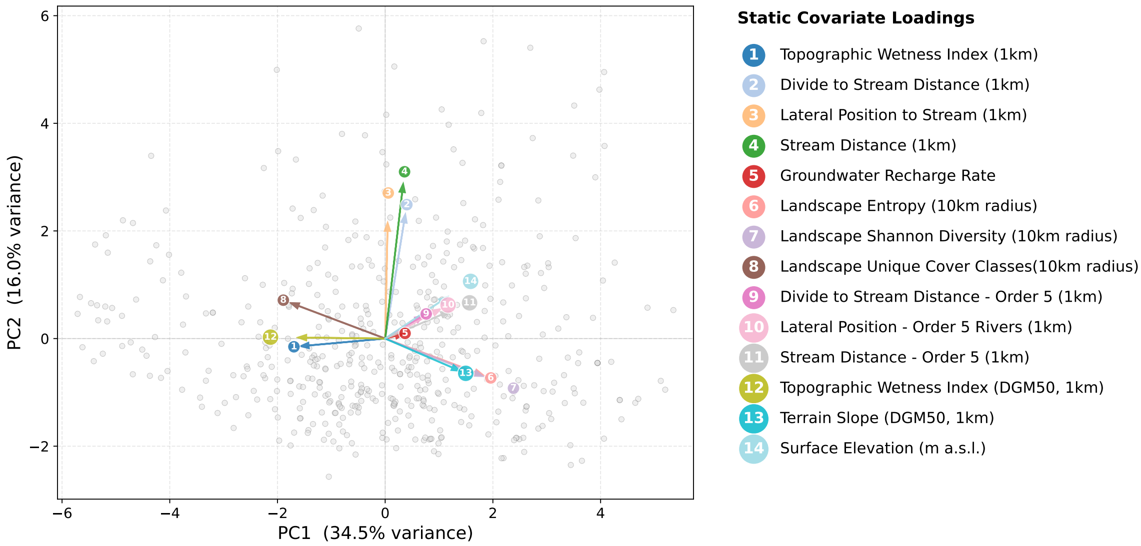

Figure 3.6.: Static Covariate PCA Components 1 & 2 Biplot. The positive PC1 direction corresponds with lower TWI and less varied landscape. The positive PC2 direction corresponds with higher stream distance.

discussed. In the remainder of this section, we decide how to split the data, accounting for the spatial and temporal aspects of the splits.

3.1.2. Data Splits

Splitting the data before modeling is essential to maintaining a realistic view of the generalization ability of any models trained. To this end, we must ensure that forecasting models cannot see the future datapoints and that the spatial models do not see the held out wells. If we let D, D_{train}, D_{val}, and D_{test} be the full, training, validation, and test sets respectively the following must hold:

$$D_{\text{train}} \cap D_{\text{val}} = D_{\text{val}} \cap D_{\text{test}} = D_{\text{train}} \cap D_{\text{test}} = \emptyset \tag{3.2}$$

Additionally, in the case of a temporal data split, the train, validation, and test sets must be contiguous and sequential in time. In this work, a sequential split on time and random split on space is used to create a valid evaluation framework for the spatiotemporal models developed in subsequent sections.

The spatiotemporal split for the data is outlined in Figure 3.8. In the time axis, 80% of the timesteps were used for the training set, 10% for the validation set, and 10% for the test set. In the space axis, 80% of the timesteps were used for the training set, 10% for the validation set, and 10% for the test set. The figure has nine sections, each determined by membership to the temporal and spatial subsets, denoted by the coding: "<temporal>-<spatial>." Several relevant sections and their purposes are expanded upon as follows. The "train-train" section denotes the dataset on which spatiotemporal models will be trained. The "val-train" section denotes the dataset belonging to the temporal validation data

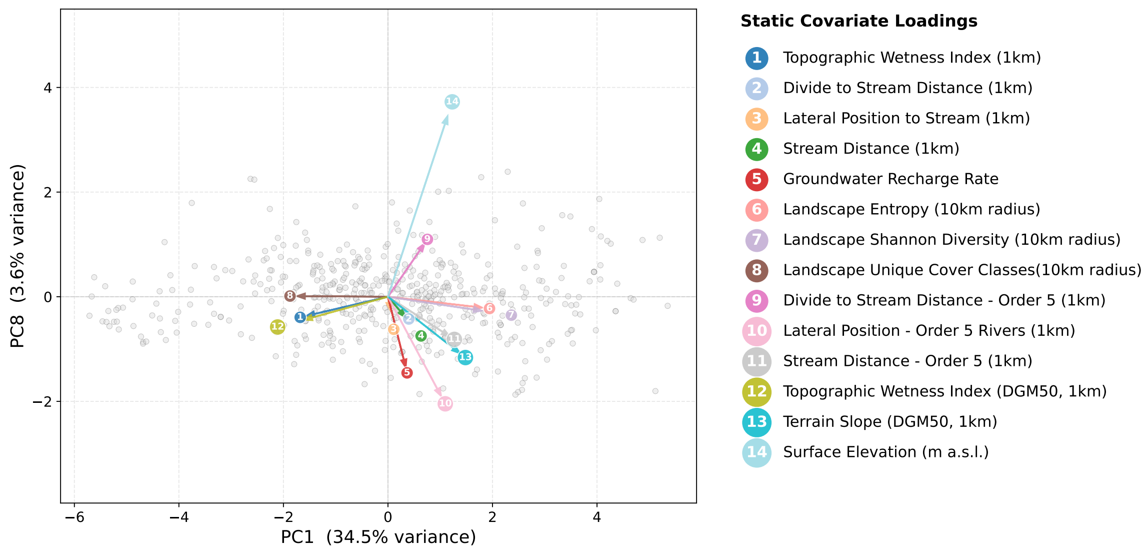

Figure 3.7.: Static Covariate PCA Components 1 & 8 Biplot. Larger values of PC8 correspond with larger elevation and smaller groundwater recharge rate.

and spatial training data, with model selection and hyperparameter optimization of the temporal models being done on this dataset. The “test-train” section denotes the dataset belonging to the temporal test data and spatial training data, and metrics computed on this set evaluate the models’ ability to do forecasting at known locations. The “val-val” dataset is primarily used in hyperparameter optimization and spatiotemporal model selection. The “test-test” section denotes the dataset belonging to the temporal test data and the spatial test data and metrics computed on this set evaluate the models’ ability to do spatiotemporal modeling at future times and unmeasured locations.

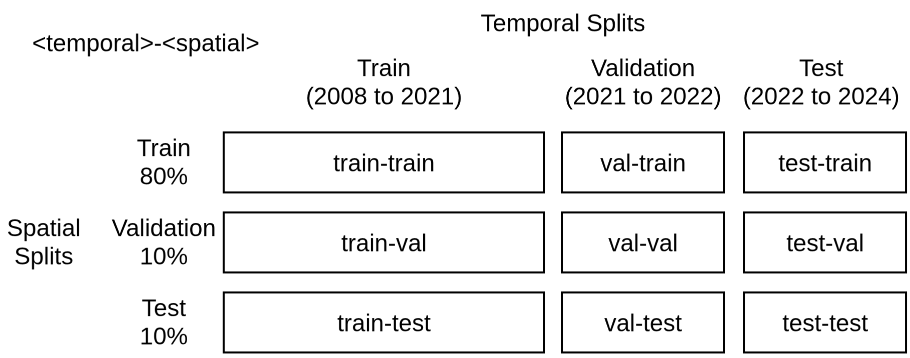

Figure 3.8.: Spatiotemporal Data Split Matrix

In Table 3.3 statistics describing the split sets are given. The time periods and number of measured timesteps is given for each of the temporal training, validation, and test sets. The number of wells and the time periods over which these wells have measurements is given for each of the spatial training, validation, and test sets.

Table 3.3.: Spatiotemporal Dataset Split Specifications

Dimension	**Steps / Counts**	**Time Period**
Temporal Split (Resolution: Weekly)		
Total duration	857 steps	2008-01-07 to 2024-06-03
Train period	684 steps	2008-01-07 to 2021-02-15
Validation period	86 steps	2021-02-22 to 2022-10-10
Test period	87 steps	2022-10-17 to 2024-06-03
Spatial Split (Disjoint IDs)		
Total unique wells	623 wells	2008-01-07 to 2024-06-03
Train wells (80%)	499 wells	2008-01-07 to 2024-06-03
Validation wells (10%)	62 wells	2008-01-07 to 2024-06-03
Test wells (10%)	62 wells	2008-01-07 to 2024-06-03

Finally, the spatial distribution within Brandenburg of the spatial training, validation, and test sets are shown in Figure 3.9. There do not appear to be large clusters of validation or test points and these points appear to be distributed throughout Brandenburg in a similar, random, manner to the training locations.

In all evaluations, windowed forecasts are made to obtain a better estimate of the true modeling error. Moreover, as is good machine learning practice, we set aside the test-* and *-test sets until the final evaluation in the results chapter, not examining the data therein before the point where all model selection decisions have been made.

3.1.3. Summary

In order to ensure the models trained are accomplishing the desired task of modeling the groundwater level in the future at unmeasured locations, the data was preprocessed and split accordingly. Moreover, data quality was ensured through addressing missing values and selecting appropriate covariates deemed relevant by subject matter experts and the literature review. Finally, the examination of scientific visualizations that uncover patterns and distributions in the data provided the groundwork for understanding what to model and the correlation structures present within the data. In Section 3.2, the approaches to modeling are expanded upon.

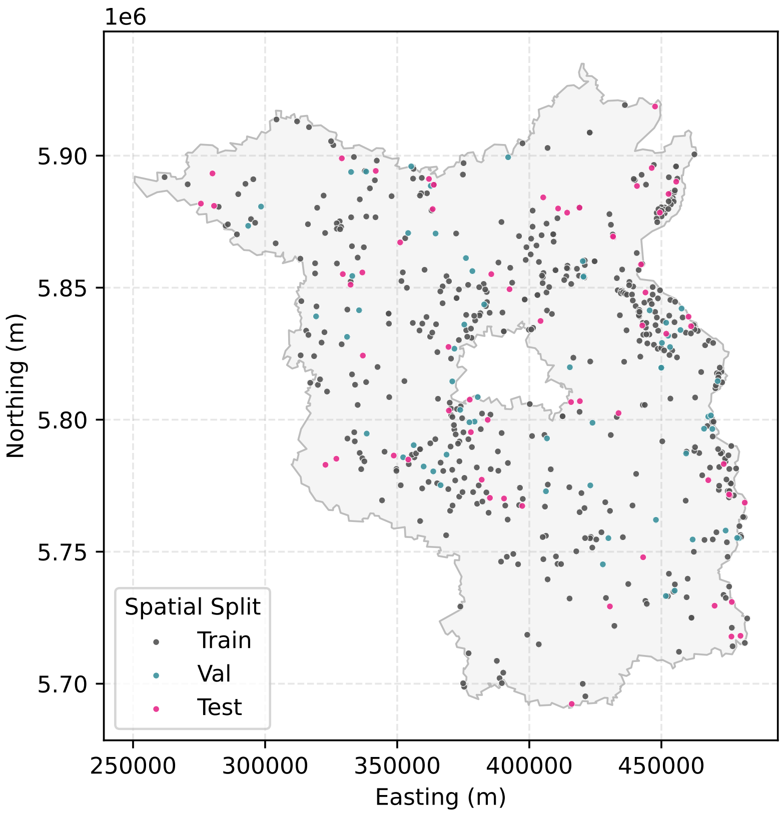

Figure 3.9.: Spatial Distribution of Split Data in Brandenburg

3.2. Modeling

This section covers the approaches for the temporal forecasting, spatial interpolation, and spatiotemporal modeling. It begins by introducing baselines and selected architectures for the temporal forecasting portion of the task. Next, it covers the baselines and selected architectures for the spatial interpolation portion of the task. Finally, it brings the two preceeding portions together and describes the approaches to the task of spatiotemporal modeling. The results are provided afterwards in Chapter 4.

3.2.1. Temporal

This section covers temporal forecasting approaches. In each of the following paragraphs, a temporal model will be described including characteristics of the model. Additionally, where applicable, the hyperparameter search space will be outlined. Before details about particular architectures are given, some aspects common amongst the models are discussed.

In all forecasting models considered, we utilize the Train-Train portion of the data splits to fit the parameters of the models. The Val-Train portion is used as a proxy for measuring the generalizability of the model in the process of selecting hyperparameters. Finally, the Test-Train portion is used to evaluate the temporal generalization of the models after all hyperparameter and design decisions are finalized. Additionally, the context given to all models is 52 previous weeks (timesteps). The forecast horizons used are one week and twelve weeks into the future. The evaluations are done with a sliding window, a stride of one timestep, and no retraining of the model for each forecast. For all temporal models except the naive forecast baseline, we predict quantiles at each forecast timestep. A justification of the models used is now given.

The temporal models implemented are: naive forecast, N-HiTS, Chronos2 zero-shot, and Chronos2 finetuned. The selection of these was driven by the following considerations. The naive forecast is utilized due to its simplicity, applicability in short-horizon forecasts, and utility in benchmarking forecasting methods [4]. The N-HiTS model is used because it has been shown in [27] to outperform other neural forecasting methods specifically on this task. The Chronos2 zero-shot model is used due to the interest in the applicability of the zero-shot capabilities of foundation models to the task of forecasting groundwater levels. The Chronos2 finetuned model is used to examine the benefits of finetuning and hyperparameter optimization on foundation models for this task. In the following paragraphs, we provide a formal definition of the loss function used in all models, provide details on the construction of the forecasting models, and discuss hyperparameter optimization where applicable.

Quantile Loss Function The quantile loss is used in all subsequent temporal forecasting algorithms which utilize a loss function. We define the quantile loss function for quantiles τ:

$$L_\tau(y, \hat{y}) = (y - \hat{y})(\tau - \mathbb{I}_{\{y<\hat{y}\}}) \tag{3.3}$$

In order to jointly train the model to optimize forecasts over a set of quantiles, $\mathcal{T}$ and N observations, the loss becomes:

$$L_{multi} = \frac{1}{N|\mathcal{T}|} \sum_{i=1}^{N} \sum_{\tau \in \mathcal{T}} L_\tau(y_i, \hat{y}_{i,\tau}) \tag{3.4}$$

Minimizing the quantile loss allows for nonparametric distributional forecasting where point estimates for a central quantile (typically 0.5) and upper and lower quantiles are made. This allows us to quantify the uncertainty in the model forecasts. In the temporal forecasting task we utilize the quantiles: $\tau \in \{0.01, 0.5, 0.99\}$.

Naive Forecast The naive forecast serves as the simplest baseline considered and makes the prediction:

$$\hat{y}_t = y_{t-1} \tag{3.5}$$

When there are timesteps farther in the future to forecast, the prediction becomes:

$$\hat{y}_t = y_{t-h} \tag{3.6}$$

with h the number of steps since the last observed value. A main assumption made is that future conditions will remain similar to past conditions. While this method is computationally inexpensive and performant with short forecast horizons, it does not account for uncertainty in forecasts or quantiles of possible futures. There are no hyperparameters to be tuned for this method and it has $\Theta(1)$ time and space complexity with respect to the size of the dataset due to a look up. For these reasons, it provides a meaningful benchmark for worst-case forecast performance based solely on historical information of the target variable.

N-HiTS N-HiTS (Neural Hierarchical Interpolation for Time Series Forecasting) is a deep learning forecasting model built on the residual stacking framework of N-BEATS, extended with hierarchical multi-rate processing [6]. As shown in Figure 3.10, the model operates at multiple temporal resolutions via hierarchical downsampling of the representation within each stack. Within each stack, a number of blocks, are composed of Multi Layer Perceptrons (MLPs) which model basis coefficients that parameterize interpolation functions used to reconstruct both backcast and forecast components. The model iteratively decomposes the signal via residual backcasting, while accumulating forecast contributions across all blocks and stacks, in effect, composing it out of learned interpolations at different sampling scales. N-HiTS scales linearly in both context and horizon as it downsamples inputs with multi-rate pooling and utilizes primarily feed-forward neural networks.

The N-HiTS model has numerous tunable hyperparameters of which we select a subset to increase comparability with the Chronos2 models. The defaults from Nixtla library (NeuralForecast) are used aside from those tuned in this thesis which are summarized in Table 3.4. We utilize the quantile loss as in Paragraph 3.2.1 and the Adam optimizer [24]. Finally, N-HiTS is considered due to the high performance on the task of groundwater level forecasting [27] as well as the competitive scaling when compared to transformer models.

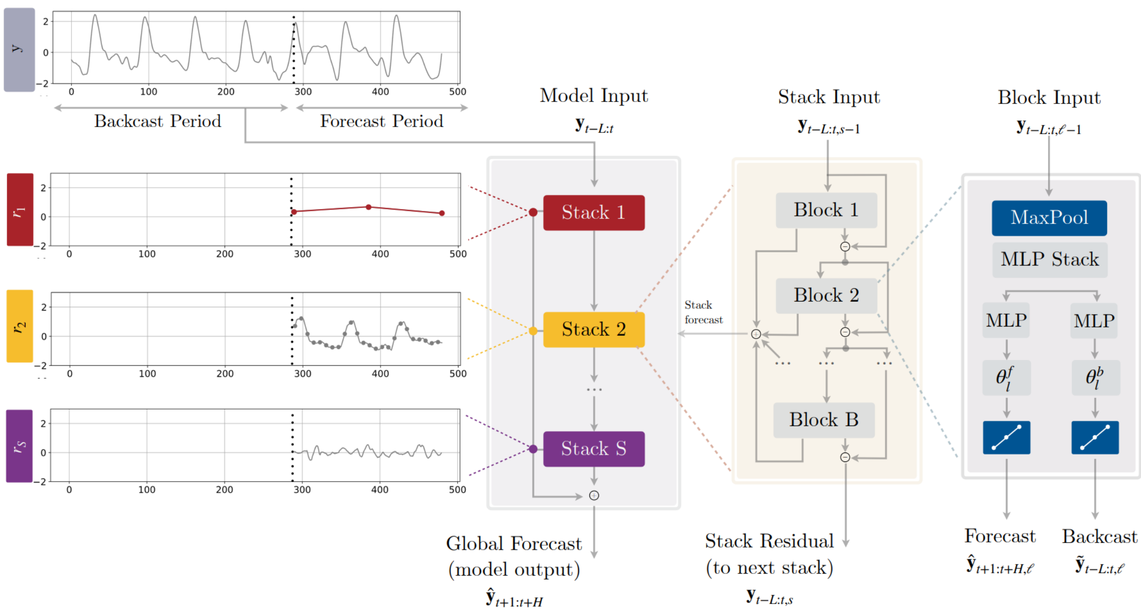

Figure 3.10.: N-HiTS Diagram. Source: Challu et al. [6]

Table 3.4.: Hyperparameter Search Space for N-HiTS

Hyperparameter	Space	Description
Max Steps	$\{500, 1000, 2000, 3000\}$	Maximum number of training iterations
Learning Rate (η)	$\left[10^{-6}, 10^{-4}\right]$	Log-scaled step size for optimization
Batch Size	$\{100, 128\}$	Count of time series samples per batch

Chronos2 Zero-Shot The Chronos2 model, introduced by Amazon [2], is a time series foundation model that allows for multivariate, covariate-informed, zero-shot quantile forecasting. The pipeline is built on a T5 transformer language model architecture, adapted to time series forecasting and operates according to Figure 3.11. It begins by robustly normalizing the inputs by applying standardization with respect to historical statistics and then taking the element-wise inverse hyperbolic sine of the values. Time indices and a mask denoting observed values are appended, the series are split into non-overlapping patches, and a residual network maps the input patches into embedding space, concluding tokenization. The transformer stack consists of the time attention which attends over the time-dimension of the input series and the group attention which attends over known covariates and other time series in a flexibly-defined group. A quantile head residual network follows to reshape the representations and map them back into target space where they are finally unscaled. Ultimately, Chronos2 scales quadratically with respect to the combined context and horizon, mitigating this transformer overhead as it compresses time steps into non-overlapping patches and decouples temporal from cross-series attention.

Chronos2 is capable of quantile forecasts and, as a foundation model, possesses zero-shot forecasting abilities. That is, given the necessary context and future covariate dataframes,

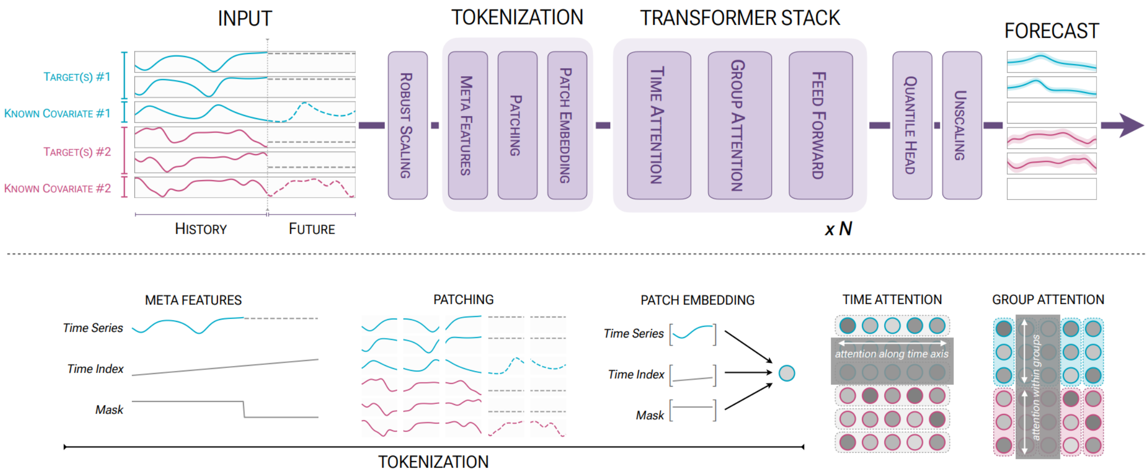

Figure 3.11.: Chronos2 Diagram. Source: Ansari et al. [2]

the model can output forecasts without adjusting internal parameters. This capability is also termed *in-context learning* and is inherited from the large language modeling and computer vision communities [5]. An advantage of zero-shot capabilities is that no training or hyperparameter optimization process is necessary to generate forecasts. In order to achieve this, models must be trained on a large corpus of example time series datasets, which in the case of Chronos2 are entirely synthetic. Chronos2 was selected to benchmark the efficacy of the zero-shot forecasting paradigm against traditional baselines and task-specific, trained models.

Chronos2 Finetuned The Chronos2 model, as a foundation model, is also capable of being finetuned, where parameter updates occur using the gradient of the multi-quantile loss as detailed in Section 3.2.1. In this case, a pre-trained Chronos2 model is optimized with the Adam optimizer [24] to minimize the loss in pursuit of better forecasting performance. In addition to adjusting all model parameters, Low-Rank Adaptation (LoRA) is considered to limit the number of parameters adjusted during fine-tuning [19]. Consequently, this architecture introduces tunable hyperparameters which are outlined in Table 3.5.

Table 3.5.: Hyperparameter Search Space for Chronos2

Hyperparameter	**Space**	**Description**
Number of Steps	$\{500, 1000, 2000, 3000\}$	Total iterations for finetuning
Learning Rate (η)	$\left[10^{-6}, 10^{-4}\right]$	Log-uniform step size for updates
Batch Size	$\{100, 128\}$	Samples processed per gradient step
Use LoRA	{True, False}	Parameter-efficient vs. full fine-tuning
LoRA r	$\{4, 8, 16, 32\}$	Rank of the low-rank update matrices
LoRA α	$\{16, 32, 64\}$	Scaling factor for LoRA layers

3.2.2. Spatial

This section covers spatial interpolation approaches, focusing on estimating groundwater levels at unmonitored geographic locations. In the following paragraphs, each spatial model is described alongside its unique structural characteristics. Where applicable, the hyperparameter search spaces and optimization strategies are outlined. Additionally, the computational complexity of each method is discussed with respect to sample size and model inputs to evaluate scaling trade-offs.

The selected spatial models range from deterministic baselines to covariate-informed Gaussian processes. The K-Nearest Neighbors (K-NN) algorithm acts as the baseline spatial interpolation method due to its simplicity, nonparametric design, and benchmarking utility. The Tabular Prior-Data Fitted Network (TabPFN) is introduced to evaluate how tabular foundation models perform on the spatial task via zero-shot in-context learning, serving as a secondary baseline that is not integrated into downstream spatiotemporal models. A residual neural network (ResNet) architecture utilizing a random Fourier feature expansion is considered to evaluate a highly scalable deep learning approach capable of resolving localized spatial variations. Finally, Gaussian Processes (GPs) are evaluated extensively as the core spatial methodology of this work due to their widespread use in hydrology and geostatistics, high expressiveness, inherent interpretability, and native probabilistic outputs.

For each of the examined models, distinct input data configurations are utilized, primarily dictated by the computational and numerical constraints encountered with exact Gaussian Processes. The baseline K-NN model is strictly restricted to geographic coordinates. TabPFN is provided with the full suite of static environmental and geologic covariates detailed in Paragraph 3.1.1. The Fourier feature ResNet is provided with the same complete covariate footprint as TabPFN. Conversely, because exact Gaussian Processes encountered severe numerical convergence and optimization stability issues when provided with raw covariates, Principal Component Analysis (PCA) was utilized to orthogonalize the input space, reduce dimensionality, and eliminate multicollinearity. Additionally, the Gaussian Processes are evaluated using reduced-dimension GeoTessera structural embeddings in place of the raw Corine land utilization fractions to maintain a compact feature domain. Ultimately, a select subset of the architectures introduced in this section serve as the spatial interpolation modules for the coupled and decoupled spatiotemporal models detailed in subsequent sections.

K-Nearest Neighbors The K-Nearest Neighbors (K-NN) algorithm is a deterministic, non-parametric, supervised learning algorithm that utilizes a neighborhood of points to predict the target value of an observation based on the distances between the points in feature space [13]. The integer hyperparameter, k, determines how many neighbors to utilize. Mathematically, for a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ and a query x, the algorithm finds the k-nearest neighborhood $N_k(x)$ by minimizing the distance metric $d(x, x_i)$ over the training samples. The regression case is utilized in this work with the prediction defined:

$$\hat{y} = \frac{1}{k} \sum_{i \in N_k(x)} y_i \tag{3.7}$$

This method, is comparatively simple with only one hyperparameter and serves as a baseline for the spatial methods tested. In terms of computational complexity, K-NN requires no explicit training phase. During inference, a brute-force implementation scales linearly with respect to the number of reference sites in the dataset, though this can be optimized to scale logarithmically. For our purposes, the K-NN algorithm is utilized only on the spatial interpolation task and is given only the coordinates of the monitoring wells during inference. Hyperparameter tuning for the optimal k is conducted and the optimal k is selected based on performance metrics on the held out validation set.

TabPFN The Tabular Prior-Data Fitted Network (TabPFN) [16, 12] is a foundational, deep learning-based meta-learner that treats supervised learning on tabular data as an In-Context Learning (ICL) problem. TabPFN passes the entire training set and test queries simultaneously through a pre-trained transformer to generate predictions. The predictions are generated with a single forward pass, eliminating the need for hyperparameter tuning.

Theoretically, TabPFN acts as a neural approximation of the Posterior Predictive Distribution (PPD) in a Bayesian framework. Given a structural prior over datasets, $P(\mathcal{D})$, generated via Structural Causal Models (SCMs) and Bayesian Neural Networks (BNNs), the network learns to approximate the intractable PPD integral for a test query x_{test} given training data $\mathcal{D}_{\text{train}}$:

$$P(y_{\text{test}}|x_{\text{test}}, \mathcal{D}_{\text{train}}) = \int P(y_{\text{test}}|x_{\text{test}}, f)P(f|\mathcal{D}_{\text{train}})df \tag{3.8}$$

By optimizing the transformer parameters Θ offline via cross-entropy loss over millions of sampled synthetic datasets, $\arg\min_{\Theta} \mathbb{E}_{\mathcal{D}\sim P(\mathcal{D})}\left[-\sum \log P_{\Theta}(y_i|x_i, \mathcal{D}_{\text{train}})\right]$, the network minimizes the expected Kullback-Leibler divergence to the true PPD, allowing it to generalize zero-shot to real-world data.

In terms of computational complexity, TabPFN requires no training or fine-tuning phase on the target dataset, evaluating predictions in a single forward pass. However, because it processes the concatenation of the training set and evaluation queries as a single sequence through self-attention layers, its computational and memory requirements scale quadratically with respect to the total number of samples. The strengths of TabPFN are high data efficiency and zero-shot inference capability. The quadratic scaling bottleneck leaves it best suited for smaller datasets; in the scope of this thesis, TabPFN was selected due to the relatively small size of the spatial dataset which exploits the high data efficiency of the model. Additionally, TabPFN without modifications is not designed for sequence modeling due to the assumption of row permutation invariance which does not hold for time series data. Due to this, TabPFN is used as a spatial baseline and provides a stationary prediction with respect to time at the inference sites. That is, no decoupled or coupled spatiotemporal models make use of TabPFN.

Spatial Fourier Feature ResNet This architecture maps localized spatial variations by combining a non-trainable feature expansion with a deep residual network. The model is designed as a deep learning baseline for spatial interpolation that can serve as a decoupled spatial embedding module for downstream spatiotemporal models. By utilizing raw geologic covariates directly, the network does not use the principal components discussed earlier, allowing the mapping layers to learn non-linear feature transformations natively during training.

Deep neural networks operating directly on low-dimensional coordinate inputs suffer from spectral bias, causing them to favor low-frequency functions and struggle to resolve high-frequency, sharp spatial variations [33]. To circumvent this, a fixed Random Fourier Feature (RFF) mapping $\gamma(\mathbf{x}) : \mathbb{R}^3 \rightarrow \mathbb{R}^{2d}$ projects the raw coordinates into a higher-dimensional hyperspace using random sinusoidal baselines [34]:

$$\gamma(\mathbf{x}) = \left[\cos(2\pi\mathbf{x}^T\mathbf{B}), \quad \sin(2\pi\mathbf{x}^T\mathbf{B})\right] \tag{3.9}$$

where elements of the frozen projection matrix $\mathbf{B} \in \mathbb{R}^{3\times d}$ are sampled from a Gaussian prior $\mathbf{B}_{ij} \sim \mathcal{N}(0, \sigma^2)$. The standard deviation σ acts as a spatial frequency tuning parameter, where higher values enable the network to resolve highly localized spatial variations.

Following the spatial expansion, the Fourier features are concatenated with the static geologic covariates $\mathbf{s}$ to construct the baseline hidden representation $\mathbf{h}_0$:

$$\mathbf{h}_0 = \mathrm{ReLU}\left(\mathbf{W}_0 \begin{bmatrix} \gamma(\mathbf{x}) \\ \mathbf{s} \end{bmatrix} + \mathbf{b}_0\right) \tag{3.10}$$

To track complex feature interactions without vanishing gradients, $\mathbf{h}_0$ passes through a sequence of Residual Blocks [44]. Each block layer l executes the mapping:

$$\begin{aligned} \mathbf{h}_{l+1} &= \mathrm{ReLU}\left(\mathbf{h}_l + \mathcal{F}(\mathbf{h}_l)\right) \\ \text{where} \quad \mathcal{F}(\mathbf{h}_l) &= \mathrm{LN}\left(\mathbf{W}_{l,2} \cdot \mathrm{Dropout}\left(\mathrm{ReLU}\left(\mathrm{LN}\left(\mathbf{W}_{l,1}\mathbf{h}_l + \mathbf{b}_{l,1}\right)\right)\right) + \mathbf{b}_{l,2}\right) \end{aligned} \tag{3.11}$$

with LN denoting Layer Normalization.

The final layer projects the deep features to output the target quantiles, providing nonparametric uncertainty intervals. Non-negativity constraints of the target variable are enforced at the output layer using a Softplus activation:

$$\hat{\mathbf{y}} = \ln\left(1 + \exp\left(\mathbf{W}_{\text{head}}\mathbf{h}_L + \mathbf{b}_{\text{head}}\right)\right) \tag{3.12}$$

The network parameters are optimized across all evaluation quantiles simultaneously using the joint multi-quantile loss function defined in Paragraph 3.2.1. Here, the target quantile set is: $\mathcal{T} = \{0.01, 0.50, 0.99\}$, providing analytical estimates for the median and extreme prediction intervals.

In terms of computational complexity, the model scales linearly with respect to both the number of training samples and the input feature dimension during training and inference. This offers a highly scalable alternative to the cubic sample scaling restrictions of exact Gaussian Processes. The model is trained using the AdamW optimizer with a weight decay of 10^{-4}. The hyperparameter search spaces are summarized in Table 3.6.

Gaussian Processes A Gaussian Process (GP) is a nonparametric, Bayesian probabilistic machine learning method that fits a distribution over functions under the assumption that any finite subset of the variables has a joint multivariate normal distribution. Theoretical

Table 3.6.: Fourier Feature ResNet Hyperparameter Space

Hyperparameter	Space	Description
Fourier Dimension (d)	$\{64, 128\}$	Number of sinusoidal projection channels.
Hidden Dimension	$[128, 512]$	Width of the intermediate residual layers.
Number of Blocks (L)	$[2, 4]$	Depth of the sequential residual block stack.
Fourier Scale (σ)	$[0.1, 30.0]$	Standard deviation of the Gaussian spatial prior.
Learning Rate (η)	$[10^{-4}, 10^{-2}]$	Log-scaled initial gradient descent step size.

equivalences exist between GPs and neural networks, as well as between GPs and Kriging variants found in geostatistics [44, 35].

Many machine learning methods, such as neural networks, utilize a parametric, weight-space view. Specifically, models take the form:

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} \tag{3.13}$$

where a prior over the weights $\mathbf{w}$ is defined through initialization, and optimization is conducted to fit the training data. GPs bypass explicit parameters by adopting a function-space view. They make no structural assumptions regarding a parametric form and instead place a prior distribution directly over the space of infinite possible functions. Mathematically, this is expressed as:

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right) \tag{3.14}$$

where $m(\mathbf{x})$ represents the mean function and $k(\mathbf{x}, \mathbf{x}')$ represents the kernel or covariance function defining the similarity between points.

A Gaussian Process is formally defined by the property that any finite collection of evaluation points must draw from a joint multivariate normal distribution. If N training points $\mathbf{X}$ and N_* test points are selected, the target values and predicted targets are jointly Gaussian:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I} & \mathbf{K}(\mathbf{X}, \mathbf{X}_*) \\ \mathbf{K}(\mathbf{X}_*, \mathbf{X}) & \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix}\right) \tag{3.15}$$

where $\sigma_n^2 \mathbf{I}$ accounts for independent, homoscedastic observation noise.

With Exact Gaussian Processes, inference is analytical. By applying Gaussian conditioning rules, the posterior distribution over predictions $\mathbf{f}_*$ is obtained directly:

$$\mathbf{f}_* \mid \mathbf{X}, \mathbf{y}, \mathbf{X}_* \sim \mathcal{N}(\boldsymbol{\mu}_*, \Sigma_*) \tag{3.16}$$

$$\boldsymbol{\mu}_* = \mathbf{K}(\mathbf{X}_*, \mathbf{X}) \left[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}\right]^{-1} \mathbf{y} \tag{3.17}$$

$$\Sigma_* = \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) - \mathbf{K}(\mathbf{X}_*, \mathbf{X}) \left[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}\right]^{-1} \mathbf{K}(\mathbf{X}, \mathbf{X}_*) \tag{3.18}$$

where $\boldsymbol{\mu}_*$ is the point prediction and Σ_* provides nonparametric uncertainty intervals for those predictions.

Training an exact GP consists of optimizing the hyperparameters of the kernel matrix itself via Type-II Maximum Likelihood Estimation (empirical Bayes), maximizing the Marginal Log-Likelihood (MLL):

$$\log p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^T \mathbf{K}_y^{-1} \mathbf{y} - \frac{1}{2} \log \left|\mathbf{K}_y\right| - \frac{N}{2} \log 2\pi \tag{3.19}$$

where $\mathbf{K}_y = \mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}$.

In terms of computational complexity, executing these exact matrix operations introduces large scaling penalties. The Cholesky decomposition required to compute the matrix inverse $\mathbf{K}_y^{-1}$ and log-determinant $\log\left|\mathbf{K}_y\right|$ scales cubically with respect to the number of training samples during the optimization phase. During inference, evaluating the predictive mean scales linearly per test query, whereas evaluating the full predictive covariance matrix scales quadratically with the training set size. Furthermore, storing the kernel matrix scales quadratically in memory space.

The kernel function chosen due to its second-order differentiability and structural relevance in modeling groundwater is the Matérn-5/2 [38]. The Matérn-5/2 kernel is defined:

$$k(x, x') = \frac{2^{-3/2}}{\Gamma(5/2)} \left(\sqrt{5}d\right)^{5/2} K_{5/2}\left(\sqrt{5}d\right) \tag{3.20}$$

where $d = \sqrt{(x - x')\Theta^{-2}(x - x')}$ is the distance scaled by the lengthscale hyperparameter and $K_{5/2}$ is a modified Bessel function. The hyperparameter search spaces are summarized in Table 3.7. To ensure stable convergence, all configurations optimize the MLL using the L-BFGS optimization algorithm with Strong-Wolfe line searches capped at 50 iterations.

Table 3.7.: Gaussian Process Hyperparameter Search Space

Hyperparameter	**Space**	**Description**
Lengthscale (ℓ)	$(0, \infty)$	Determines how quickly the covariance decays as distance between coordinates increases.
Noise Variance (σ_n^2)	$(0, \infty)$	Accounts for measurement error.

Because specific spatial and environmental inputs directly alter the dimensionality and additive structure of the covariance matrix, several architectural variants of Gaussian Processes are examined in the following paragraphs.

Gaussian Process on Coordinates The baseline spatial model maps geographic spatial correlation using strictly coordinate pairs ($\mathbf{x} \in \mathbb{R}^2$). The architecture utilizes a single constant mean module and a scaled Matérn-5/2 covariance module configured with Automatic Relevance Determination (ARD). Enabling ARD over the two spatial dimensions allows the model to learn independent lengthscale hyperparameters for the x and y axes. This accomodates anisotropy and allows for feature relevance determination using the inverse lengthscale for each feature. The model scales cubically with respect to the number of training stations due to the dense spatial Cholesky factorizations required at each optimization step.

Gaussian Process on Coordinates and Geologic Principal Components This variant expands the input space by appending geologic principal components to the geographic coordinate vectors ($\mathbf{x} \in \mathbb{R}^{2+d_{\text{env}}}$). Instead of a single multi-dimensional Matérn-5/2 kernel, this architecture implements a decoupled additive covariance structure: $\mathbf{K} = \mathbf{K}_{\text{space}} + \mathbf{K}_{\text{env}}$. By isolating dimensions, $\mathbf{K}_{\text{space}}$ operates strictly on the physical coordinates with ARD, while $\mathbf{K}_{\text{env}}$ handles the geological principal components. This additive composition allows for interpretability and limits convergence issues. Computational complexity remains cubically bound to sample size, introducing only linear time overhead relative to the total number of features during pairwise distance matrix operations.

Gaussian Process on Coordinates and Reduced Tessera Embeddings This variant expands the input space by appending the reduced Tessera embeddings to the geographic coordinates. It also utilizes an additive covariance structure $\mathbf{K} = \mathbf{K}_{\text{space}} + \mathbf{K}_{\text{tessera}}$ and utilizes ARD. The computational complexity is as above and it serves to determine how the reduced Tessera embeddings alter the Gaussian process performance without the geologic principal components.

Gaussian Process on Coordinates, Geologic Principal Components, and Reduced Tessera Embeddings The final variant combines the geographic coordinates, the geologic principal components, and the reduced Tessera embeddings. It does this by, allocating an independent, ARD-enabled Matérn-5/2 kernel and scale module to each group: $\mathbf{K} = \mathbf{K}_{\text{space}} + \mathbf{K}_{\text{env}} + \mathbf{K}_{\text{tessera}}$. Computational complexity again remains cubically bound to sample size, introducing only linear time overhead relative to the total number of features during pairwise distance matrix operations.

3.2.3. Decoupled Spatiotemporal

This section is the first full spatiotemporal section covering models with decoupled temporal and spatial components. The decoupled approach is where the forecasting and spatial interpolation are done in two separate steps, as in Figure 3.12. First, a forecast model makes predictions at sites where the groundwater level history is known. Then a Gaussian Process uses these forecasts as labels and makes predictions on the sites where the groundwater level is unknown. In effect, this produces forecasts at unmonitored sites. Mathematically we model:

$$P(y_{t:t+H}^{\text{unmonitored}} | \hat{y}_{t:t+H}^{\text{monitored}}, X), \text{ with } \hat{y}_{t:t+H}^{\text{monitored}} = \hat{f}(y_{t-C:t}, X) \tag{3.21}$$

The result is that the combination of a spatial model and a temporal model allow for spatiotemporal modeling and forecasts at sites where only static geological features are known.

The spatial model used is the Gaussian Process due to its ability to interpolate values and widespread utilization in the geosciences. The forecast models used in this section are all those utilized in Section 3.2.1. The analyses are conducted on the Val-Val and Test-Test sets. In order to accomplish this, the mentioned forecast models forecast the values at the

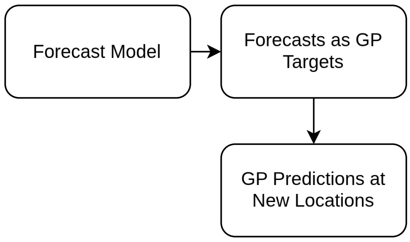

Figure 3.12.: Decoupled Pipeline Diagram

Val-Train and Test-Train sets. The Gaussian Processes use the forecasted values to obtain the interpolated forecast at locations with no historical groundwater levels.

In all of the following model variants, the forecasts at horizons of 1 and 12 weeks for each date are used as targets for the Gaussian Processes with the setup of each Gaussian Process the same as in Section 3.2.2. That is there are individual Gaussian Process models trained for each timestep, horizon, forecast model, and feature set. The differences between each of the following model variants is which features are input to the kernel as discussed below.

The kernels used for each of the Gaussian processes on forecasts mirror those used in the spatial-only section. Specifically, the **Gaussian Processes on Coordinates** uses a single Matérn-5/2 kernel with the x and y coordinates given. The **Gaussian Processes on Coordinates and Geologic Principal Components** uses a sum of two Matérn-5/2 kernels, one for coordinates and one for geologic principal components. The **Gaussian Processes on Coordinates, Geologic Principal Components, and Tessera Principal Components** uses a sum of three Matérn-5/2 kernels, one for coordinates, one for geologic principal components, and one for tessera principal components. Finally, the Gaussian Process predicts standard deviations which we use to create 98% Bayesian Credible intervals for forecast intervals.

3.2.4. Coupled Spatiotemporal

The coupled spatiotemporal models do not make direct use of the forecast model predictions as targets for a spatial model. Instead the models are trained to use the static covariates of each site and the histories of the monitored sites to predict the future groundwater levels at the unmonitored sites directly. Mathematically we model:

$$P(y_{\text{unmonitored}}|X, y_{historical}) \tag{3.22}$$

In contrast to the multi-step framework detailed in Section 3.2.3, the coupled spatiotemporal paradigm rejects the formulation where a spatial model is strictly downstream from independent temporal predictions. Instead of treating temporal forecasts as intermediate targets or labels for an isolated spatial interpolation step, coupled architectures unify these domains into a single, end-to-end mapping function. The models in this section ingest the static geological covariates of each site alongside the historical time series of monitored locations simultaneously to predict the target groundwater levels at unmonitored locations directly.

This section evaluates three primary variations of this coupled design. First, a computationally optimized Kronecker Gaussian Process is presented to model dense spatiotemporal data grids under structured conditions. Second, the specialized end-to-end mini-batch training routine utilized to optimize the neural network variants is detailed. Third, a pure Cross-Attention model which uses multi-head attention layers and a deterministic multi-quantile output head to generate distributional forecasts is shown. Finally, we show a hybrid Gaussian Process with Cross Attention which utilizes the attention stack as a structured prior mean function, delegating residual uncertainty to a multi-task Sparse Variational GP.

In these attention models, we utilize temporal and spatial modules to embed information about the spatial and temporal dynamics of the system before combining this information with the attention mechanism. The temporal forecasting backbone is instantiated using the frozen N-HiTS architecture described in Section 3.2.1. Spatial representation extraction relies on the frozen ResNet backbone established in Section 3.2.2.

Kronecker Gaussian Processes While the aforementioned exact Gaussian Processes (GPs) are infeasible in the naive implementation when there are many data points due to the $O(M^2)$ space complexity from the materialization of the covariance matrix in memory (with M the number of observations), there is a viable alternative when the dataset exhibits structure expanded upon in this section. This alternative allows us to fit a GP to the Train-Train set and do inference on the Val-Val (and Test-Test) set without exhausting our computational resources. Specifically, when data are collected on a complete spatiotemporal Cartesian grid (not sparse), ($\mathbf{X} = \mathbf{X_s} \times \mathbf{X_t}$, where $M = NT$), the joint training covariance matrix exhibits a Kronecker structure ($\mathbf{K}_{cc} = \mathbf{K}_s \otimes \mathbf{K}_t$).

While the full derivation is expanded upon in Saatci [37], we present the mathematical derivation of our application. To evaluate the predictive weight linear system $\boldsymbol{\alpha} = (\mathbf{K}_{cc} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y}$ without constructing the dense $NT \times NT$ joint covariance matrix, this method leverages a simultaneous spectral diagonalization which exploits the Kronecker structure of the data. This approach converts the global inversion problem into cheap, independent operations on the spatial and temporal components, reducing computational complexity from $O(N^3T^3)$ to $O(N^3 + T^3)$. The implementation proceeds with four steps:

Component Eigendecomposition: The independent spatial ($\mathbf{K}_s \in \mathbb{R}^{N\times N}$) and temporal ($\mathbf{K}_t \in \mathbb{R}^{T\times T}$) training covariance matrices are decomposed into their respective orthonormal eigenspaces:

$$\mathbf{K}_s = \mathbf{Q}_s\boldsymbol{\Lambda}_s\mathbf{Q}_s^\top, \quad \mathbf{K}_t = \mathbf{Q}_t\boldsymbol{\Lambda}_t\mathbf{Q}_t^\top \tag{3.23}$$

where $\mathbf{Q}_s, \mathbf{Q}_t$ represent orthogonal eigenvector matrices, and $\boldsymbol{\lambda}_s = \mathrm{diag}(\boldsymbol{\Lambda}_s)$, $\boldsymbol{\lambda}_t = \mathrm{diag}(\boldsymbol{\Lambda}_t)$ are the isolated eigenvalue vectors.

Joint Eigenspace Construction: The eigenvalues of the joint Kronecker system ($\mathbf{K}_s \otimes \mathbf{K}_t + \sigma_n^2\mathbf{I}$) are mapped directly onto a compact 2D grid matrix $\mathbf{L}_{\text{noisy}} \in \mathbb{R}^{N\times T}$ via an outer product, bypassed by a stability clamp to guarantee numerical invertibility:

$$\mathbf{L}_{\text{noisy}} = \max\left(\boldsymbol{\lambda}_s\boldsymbol{\lambda}_t^\top + \sigma_n^2\mathbf{1}\mathbf{1}^\top,\ \epsilon\mathbf{1}\mathbf{1}^\top\right) \tag{3.24}$$

Target Coordinate Rotation: The row-major flattened target vector $\mathbf{y}$ is reshaped to its original grid dimension $\mathbf{Y} \in \mathbb{R}^{N\times T}$ and projected into the joint eigenspace. Utilizing row-major Kronecker multiplication identities, this step bypasses Kronecker vectorization in favor of highly optimized matrix multiplications:

$$\mathbf{Y}_{\text{rot}} = \mathbf{Q}_s^\top \mathbf{Y} \mathbf{Q}_t \tag{3.25}$$

Diagonal Inversion and Back-Projection: Because the joint operator is strictly diagonalized within this rotated coordinate system, inversion reduces to element-wise Hadamard division (⊘) by the noisy joint eigenvalue matrix. The result is transformed back to the original spatiotemporal coordinate system and flattened to yield the exact solution vector $\boldsymbol{\alpha}$:

$$\mathbf{A} = \mathbf{Q}_s \left(\mathbf{Y}_{\text{rot}} \oslash \mathbf{L}_{\text{noisy}}\right) \mathbf{Q}_t^\top, \quad \boldsymbol{\alpha} = \text{vec}_R(\mathbf{A}) \tag{3.26}$$

For a full derivation, consult Appendix C.1 and the mentioned source [37]. The result is that we obtain a function of the form:

$$f(x, y, t, c) = \hat{\text{gwl}} \tag{3.27}$$

where x, y are geographic coordinates, t is time, and c represents covariates. With this function, we are able to do spatiotemporal modeling, though this approach, for simplicity and due to its feasibility, does not utilize forecast horizons.

Neural Network Training Procedure In order to train the spatiotemporal neural network variants end-to-end, a specialized spatial partitioning optimization procedure is required to enforce simultaneous spatial interpolation and temporal alignment. Let $\mathcal{W}$ denote the complete universe of available well identifiers within the training set, where each well $w \in \mathcal{W}$ is associated with time series $\mathcal{Y}_w$ and a static covariate vector $\mathcal{X}_w$. At the beginning of each training epoch, the system dynamically partitions the global set into two disjoint subsets based on a target holdout ratio $r \in (0, 1)$. This stochastic assignment splits the entities such that $\mathcal{W} = \mathcal{W}_{\text{context}} \cup \mathcal{W}_{\text{target}}$, where the holdout target set satisfies $|\mathcal{W}_{\text{target}}| = \lfloor r \times |\mathcal{W}| \rfloor$.

The optimization pipeline proceeds by evaluating the base temporal forecasting network exclusively over the context domain, mapping the historical observations to a predictive horizon matrix $\mathbf{V}' = \text{NHITS}(\mathcal{Y}_{w \in \mathcal{W}_{\text{context}}}) \in \mathbb{R}^{|\mathcal{W}_{\text{context}}| \times h}$. Concurrently, the spatial representations for both domains are projected through the key-query embedder to yield context keys $\mathbf{K} \in \mathbb{R}^{|\mathcal{W}_{\text{context}}| \times \text{emb\_dim}}$ and target queries $\mathbf{Q} \in \mathbb{R}^{|\mathcal{W}_{\text{target}}| \times \text{emb\_dim}}$. The multi-head cross-attention mechanism uses these representations to compute a context-aware target matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{W}_{\text{target}}| \times \text{emb\_dim}}$, exposing the target wells to the predictive histories of the surrounding context wells weighted by their geological similarity.

The final loss calculation is evaluated strictly on the unmonitored target sites by comparing the head predictions against the true forward time series $\mathbf{Y}_{\text{true}} = \mathcal{Y}_{w \in \mathcal{W}_{\text{target}}} \in \mathbb{R}^{|\mathcal{W}_{\text{target}}| \times h}$. For the attention-only variant, parameters are optimized directly via backpropagation of the multi-quantile pinball loss $\mathcal{L}_{\text{MQL}}(\Phi_o(\mathbf{A}), \mathbf{Y}_{\text{true}})$. For the attention-GP variant, the network updates the variational parameters and embedder weights by maximizing the evidence lower bound $\mathcal{L}_{\text{ELBO}}$ evaluated on the target residual matrix $\tilde{\mathbf{Y}} = \mathbf{Y}_{\text{true}} - \Phi_{\text{mean}}(\mathbf{A})$. By shifting the holdout compositions $\mathcal{W}_{\text{target}}$ across successive epochs, the model is trained without being able to memorize static spatial configurations, forcing the network to learn generalized spatiotemporal interpolation-forecasting dynamics. All neural models in this section utilize the Adam optimizer for training.

Cross Attention The cross attention model seeks to utilize the best performing temporal forecaster on the validation set and the ResNet spatial model to create a powerful spatiotemporal model. To do this, the attention mechanism used in modern transformer architectures, as shown in Vaswani et al. [42], is used with the values being context forecasts, the keys being the context spatial embeddings and the queries being the target location spatial embeddings. The attention mechanism then attends to the relevant parts of the forecasts of known, context wells based on the similarity between the static covariates of the context and target wells in the embedding space.

The cross attention model uses the history of the context wells and makes forecasts using a frozen N-HiTS model. The ResNet embedder from Paragraph 3.2.2 is frozen and used to generate embeddings from the static covariates at both context and target locations by extracting the hidden representation of the inputs just before the output layer. The cross attention module uses a final feed forward layer to transform the embeddings and forecasts before fusing them with the attention equation. The final output is a tensor of distributional forecasts over the specified horizon at the target well sites. The model was implemented using pytorch and trained to minimize the multi quantile loss as shown previously.

It is described mathematically as follows. Let $V' \in \mathbb{R}^{|\mathcal{W}_{\text{context}}| \times h}$ denote the tensor of N-HiTS median forecasts across $|\mathcal{W}_{\text{context}}|$ context wells over a forecast horizon h. This tensor is mapped to a latent space via a value embedder, $\Phi_v : \mathbb{R}^{|\mathcal{W}_{\text{context}}| \times h} \longrightarrow \mathbb{R}^{|\mathcal{W}_{\text{context}}| \times \text{emb\_dim}}$, yielding the attention values matrix V:

$$V = \Phi_v(V') \in \mathbb{R}^{|\mathcal{W}_{\text{context}}| \times \text{emb\_dim}} \tag{3.28}$$

where emb_dim denotes the embedding dimension. The keys $K \in \mathbb{R}^{|\mathcal{W}_{\text{context}}| \times \text{emb\_dim}}$ and queries $Q \in \mathbb{R}^{|\mathcal{W}_{\text{target}}| \times \text{emb\_dim}}$ are derived by applying a shared key-query embedder, Φ_{kq}, to the raw spatial embeddings of the context and target wells, respectively. Using these representations, the scaled dot-product attention mechanism with a residual connection is formalized as:

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{\text{emb\_dim}}}\right) V + Q \tag{3.29}$$

where $A \in \mathbb{R}^{|\mathcal{W}_{\text{target}}| \times \text{emb\_dim}}$ represents the updated target well representations. Finally, the context-aware representation matrix A is projected through an output head, $\Phi_o : \mathbb{R}^{|\mathcal{W}_{\text{target}}| \times \text{emb\_dim}} \longrightarrow \mathbb{R}^{|\mathcal{W}_{\text{target}}| \times h \times q}$. The resulting tensor provides multi-quantile forecasts (q) across the temporal horizon h for each of the $|\mathcal{W}_{\text{target}}|$ target wells. A diagram of the architecture is depicted in Figure 3.13.

There are several hyperparameters to tune which pertain primarily to the cross attention module. These hyperparameters are specified in Table 3.8 and were optimized using optuna with 30 trials.

Gaussian Process with Cross Attention The spatiotemporal cross-attention model is extended by integrating a Sparse Variational Gaussian Process (SVGP) head directly onto the latent representations. Rather than relying on a deterministic linear projection head to generate quantile intervals, this architecture treats the cross-attention output as a highly

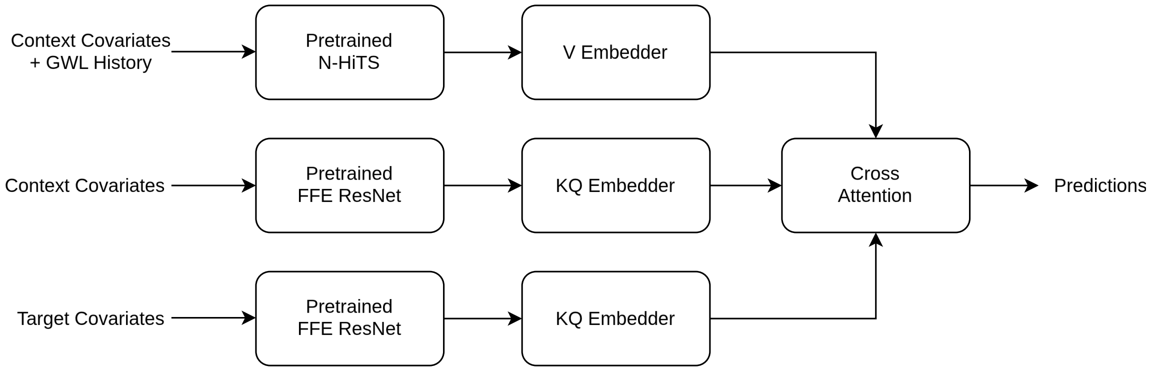

Figure 3.13.: Cross Attention Architecture Diagram

Table 3.8.: Cross-Attention Hyperparameter Search Space

Hyperparameter	**Space**	**Description**
Embedding Dimension	$\{32, 64, 128, 256\}$	Dimensionality of the embedding space for the cross-attention layers.
Number of Heads	$\{1, 2, 4, 8, 16\}$	Number of parallel attention heads.
Attention Layers	$[1, 4]$	Total count of cross-attention blocks stacked sequentially.
Learning Rate (lr_{new})	$[10^{-4}, 10^{-2}]$	Step size parameter for the new attention layers, optimized on a logarithmic scale.
Dropout	$[0.0, 0.4]$	Regularization probability applied within the cross-attention blocks.

structured, data-driven prior mean function, while a multi-task SVGP captures localized spatiotemporal covariance structures in the residuals.

Crucially, the GP kernel does not operate within the physical coordinate space ($\mathbb{R}^2$ or $\mathbb{R}^3$), nor does it require an independent latent encoder. Resorting to this formulation, the kernel evaluation occurs directly within the learned, bounded latent space produced by the key-query embedder, Φ_{kq}. Consequently, the identical geological and spatial features that govern the multi-head cross-attention weighting mechanism simultaneously dictate kernel covariance. If two wells are geographically separated but exhibit strong geologic similarity, the shared embedding projection forces them closer together within the kernel space, naturally driving higher predictive covariance.

To bypass the typical $O(N^3)$ computational bottleneck associated with standard Gaussian processes over a large spatiotemporal dataset of well observations, and to allow gradient flow to the attention head while using Gpytorch, a sparse approximation via a set of M learnable inducing points $Z_u \in \mathbb{R}^{M \times \text{emb\_dim}}$ is introduced. The model is trained end-to-end to maximize the Evidence Lower Bound (ELBO), formulated as:

$$\mathcal{L}_{\text{ELBO}} = \sum_{t=1}^{T} \mathbb{E}_{q(f_t)} \left[\log p(\mathbf{y}_t \mid f_t)\right] - D_{\text{KL}}\big(q(u) \parallel p(u)\big) \tag{3.30}$$

where $q(u) \sim \mathcal{N}(\mathbf{m}, \mathbf{S})$ is the variational distribution over the inducing variables, parameterized using a lower-triangular Cholesky factor. Under this objective, the target values fed into the variational log-likelihood are explicit residuals computed by subtracting the attention-derived mean from the ground truth targets, $\tilde{\mathbf{y}} = \mathbf{y} - \mu_{\text{attention}}$. By minimizing the ELBO with respect to these residuals, the latent GP prior is formulated with a mean of zero, shifting the burden of capturing the primary global trend onto the attention network. Gradients propagate end-to-end through the ELBO, indirectly guiding the attention mean head to capture regional, macroscopic trends while encouraging the SVGP to correct localized discrepancies.

Mathematically, the formulation builds directly upon the attention-derived representation matrix $A \in \mathbb{R}^{|\mathcal{W}_{\text{target}}| \times \text{emb\_dim}}$. The structural mean is obtained via a linear projection layer:

$$\mu_{\text{attention}} = \Phi_{\text{mean}}(A) \in \mathbb{R}^{|\mathcal{W}_{\text{target}}| \times h} \tag{3.31}$$

Simultaneously, the queries $Q \in \mathbb{R}^{|\mathcal{W}_{\text{target}}| \times \text{emb\_dim}}$ are passed directly to an independent multi-task GP model, which instantiates h parallel latent tasks to mirror the forecast horizon. For each step along the horizon, the latent stochastic process maps the shared queries into a zero-centered residual space, evaluating spatial covariance using a Matérn-5/2 kernel with Automatic Relevance Determination (ARD):

$$f(Q) \sim \mathcal{GP}\left(\mathbf{0}, \mathcal{K}_{\text{Matérn}}(Q, Q')\right) \tag{3.32}$$

The joint spatiotemporal predictive distribution for the target wells is subsequently derived through a multi-task Gaussian likelihood that incorporates per-horizon observation noise:

$$p(\mathbf{y} \mid Q) = \mathcal{N}\left(\mu_{\text{attention}} + \mu_{\text{GP}}, \Sigma_{\text{GP}} + \sigma^2_{\text{likelihood}}\mathbf{I}\right) \tag{3.33}$$

A diagram of this joint architecture is depicted in Figure 3.14.

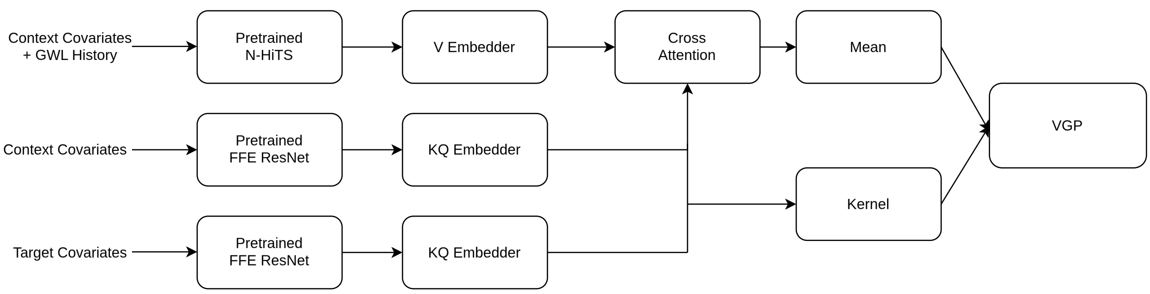

Figure 3.14.: Gaussian Process Cross Attention Architecture Diagram

The hyperparameter search space for this spatiotemporal model includes both attention parameters and specific variational GP parameters. Due to the high modeling capacity of the SVGP head, the attention layers are kept shallow to preserve a clean residual signal. These parameters are specified in Table 3.9 and were optimized using Optuna with 30 evaluation trials.

Table 3.9.: Attention-GP Hyperparameter Search Space

Hyperparameter	Space	Description
Embedding Dimension	$\{32, 64, 128\}$	Dimensionality of the joint embedding layer used within the attention mechanism.
Number of Heads	$\{1, 2, 4, 8\}$	Number of parallel attention heads.
Attention Layers	$[1, 2]$	Total count of attention blocks stacked sequentially; kept shallow to ensure a cleaner Gaussian Process residual.
Learning Rate (lr)	$[10^{-4}, 10^{-2}]$	Initial step size parameter optimized on a logarithmic scale for the gradient descent algorithm.
Dropout	$[0.0, 0.4]$	Regularization probability applied within the attention layers to prevent overfitting.
Inducing Points (n_{inducing})	$\{32, 64, 128\}$	Number of latent inducing variables optimized for the sparse Gaussian Process approximation.

3.3. Experimental Setup

In this section, we expand upon the metrics and evaluation process in order to define how we will measure and compare performances of the chosen models on the data. The metrics section contains metrics to evaluate the forecast performance, the deviation from the targets, and the calibration, sharpness, and coverage of the prediction intervals. Additionally, there are some metrics from the HydroML literature. In the evaluation procedure subsection, discussion of the process for evaluating the models is covered. Specifically, the dataset splits and metrics used, and other details of the computation of metrics are outlined.

3.3.1. Metrics

The metrics chosen for evaluation were decided upon based on their relevance to the problem and relevance to the domain. The Root Mean Squared error, for example, is a widely used metric in regression machine learning tasks with the benefit that it has the same units as the target variable. Mean Absolute Error is another metric for measuring average deviation from the target though it does not scale outliers as squared errors do. The Nash-Sutcliffe Efficiency and Kling-Gupta Efficiency, for example, are popular in hydrology and specifically in forecasting problems within hydrology. Finally, three metrics to assess the intervals of forecasts and spatiotemporal predictions are included, the Mean Prediction Interval Width, the Winkler Score, and the Prediction Interval Coverage Probability. These allow the user to identify forecasts and predictions that are over-/under-confident as well as examine the coverage of the forecast and prediction intervals.

Additionally, the data on which the metrics are calculated are described in the core-sponding sections of Chapter 4. The evaluation of the spatial generalization, temporal forecasting, and spatiotemporal modeling performances must be treated differently as they are different problems.

Root Mean Squared Error (RMSE) The Root Mean Squared Error measures the orthogonal distance between observed values y and predictions $\hat{y}$ within the target space and has the same units as the target. For a sample size of N, it is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{3.34}$$

Nash-Sutcliffe Efficiency (NSE) The Nash-Sutcliffe Efficiency is a metric, identical to the R^2 (coefficient of determination), used in hydrology for evaluating forecasting performance. It is defined as follows:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{3.35}$$

where $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ represents the empirical mean of the observed values.

Kling-Gupta Efficiency (KGE) The Kling-Gupta Efficiency is a holistic diagnostic metric that decomposes predictive performance into three structurally independent components: correlation, variability, and bias. It is expressed as:

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{3.36}$$

with the constituent terms:
Linear Correlation Component (r): The Pearson correlation coefficient between observed and predicted values:

$$r = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{3.37}$$

Variability Ratio Component (α): The ratio of the standard deviation of the predictions to the observed standard deviation:

$$\alpha = \frac{s_{\hat{y}}}{s_y} \tag{3.38}$$

Bias Component (β): The ratio of the mean prediction to the observed mean:

$$\beta = \frac{\bar{\hat{y}}}{\bar{y}} \tag{3.39}$$

Mean Absolute Error (MAE) The Mean Absolute Error quantifies the average linear magnitude of residuals between observed values y and predictions $\hat{y}$. Unlike the RMSE, it weights all individual errors equally without penalizing larger outliers disproportionately, providing a robust baseline of average model error in the target units. This metric is utilized exclusively within the spatial validation framework where the decompositions of the Kling-Gupta Efficiency (KGE) become mathematically ineffective or unstable. For a sample size of N, it is defined as:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{3.40}$$

Mean Prediction Interval Width (MPIW) The Mean Prediction Interval Width assesses the average sharpness and resolution of a probabilistic or quantile-based prediction interval. For an upper quantile bound $\hat{y}_{1-\frac{\alpha}{2}}$ and a lower quantile bound $\hat{y}_{\frac{\alpha}{2}}$ defining a $(1-\alpha)$ nominal confidence coverage, the metric evaluates across N instances as:

$$\mathrm{MPIW} = \frac{1}{N}\sum_{i=1}^{N} \left(\hat{y}_{i,1-\frac{\alpha}{2}} - \hat{y}_{i,\frac{\alpha}{2}}\right) \tag{3.41}$$

Winkler Score The Winkler Score evaluates interval forecasts by simultaneously assessing the sharpness (width) of the prediction interval and penalizing empirical target coverage failures (excursions beyond the specified bounds). Given a nominal significance level α, which defines a $(1-\alpha)$ nominal confidence coverage, let $L_i = \hat{y}_{i,\frac{\alpha}{2}}$ represent the lower predictive bound and $U_i = \hat{y}_{i,1-\frac{\alpha}{2}}$ represent the upper predictive bound for observation y_i. The Winkler Score for a single observation y_i is defined as:

$$W_\alpha(L_i, U_i, y_i) = (U_i - L_i) + \frac{2}{\alpha}(L_i - y_i)\mathbb{I}(y_i < L_i) + \frac{2}{\alpha}(y_i - U_i)\mathbb{I}(y_i > U_i) \tag{3.42}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

Prediction Interval Coverage Probability (PICP) The Prediction Interval Coverage Probability (PICP) measures the empirical reliability of interval forecasts by calculating the proportion of true target values that successfully fall within the predicted lower and upper bounds. Let L_i represent the lower predictive bound and U_i represent the upper predictive bound for observation y_i across a dataset of N samples. It is defined as:

$$PICP = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(L_i \leq y_i \leq U_i) \tag{3.43}$$

For a well-calibrated model, the PICP should closely approximate the nominal confidence level $1-\alpha$.

3.3.2. Evaluation Procedure

Temporal The temporal models were trained on the Train-Train, validated on the Val-Train, tested on the Test-Train sets. In order to reduce the error in the estimation of the metrics, sliding window evaluation of the models is done. That is, for each step where there are sufficient future true labels, a forecast is made, then the time index is incremented and another forecast is made. The metrics utilized are RMSE, KGE, NSE, MPIW, PICP, and Winkler.

Spatial The spatial models are trained on the Train-Train, validated on the Val-Val, and tested on the Test-Test sets so they can be more easily compared with the spatiotemporal models in future sections, especially in the relative comparisons section. It is additionally noteworthy that the temporal component of the evaluation data is "averaged-out" in that the spatial models make one prediction per location and this is compared to the mean groundwater level over time at that location. This allows for examination of the spatial generalization with less weight placed on the ability of the model to match the variation of a given site. The metrics utilized are spatial RMSE, spatial NSE, spatial MAE, MPIW, PICP, and Winkler. The KGE is omitted since there is no longer variability in the predictions at a given site. Additionally, the "spatial-" prefix indicates the metrics are computed over the set of locations with residuals of the form: $e_s = \bar{y}_s - \hat{y}_s$ with s denoting the site.

Spatiotemporal The spatiotemporal models (coupled and decoupled) are trained on Train-Train, validated on Val-Val, and tested on Test-Test so that they can be evaluated for their ability to forecast at sites with no historical groundwater level timeseries. That is, spatiotemporal forecasts are made using the given context with stride one and then the evaluation proceeds as for temporal models at the sites where no historical context was used. The metrics utilized are RMSE, NSE, KGE, MPIW, PICP, and Winkler.

4. Results

This section reports the results of the data and methodology discussed in the preceeding chapter. It begins with temporal modeling results, then moves onto spatial modeling results, and finally the spatiotemporal modeling results. In each section, aspects such as the data used, the hyperparameters selected for each model, and the values for each of the evaluation metrics are reported. Additionally, visualizations are created to allow for comparison between methods. The section closes with specific analyses and relative comparisons of a decoupled spatiotemporal model.

4.1. Temporal

The data splits used in the following temporal analyses are the Train-Train, Val-Train, and Test-Train. The focus is on the temporal forecasting ability of the monitoring wells but there should not be leakage from the spatial validation or test sets to influence modeling decisions. This helps to mitigate the risk of developing models that overfit the entire dataset and generalize poorly.

In each of the subsections, the selected hyperparameters are reported with median validation set RMSE, MPIW, and PICP over sites. To maintain consistency across horizons, the neural models were configured to output a 12-week (h12) forecast vector simultaneously; the one-week horizon (h1) evaluations were extracted directly from the first step of this multi-horizon prediction. These results are reported for both forecast horizons of one week and twelve weeks and aggregated over the monitoring wells in the train spatial split. In effect, the metrics aggregate the performance of the forecast at the given horizon and sites, measuring the overall forecasting ability at all sites considered. In the final subsection, aggregated results with additional metrics computed on the test set are visualized to allow for better comparison between models.

4.1.1. Naive Forecast

The naive forecast has no hyperparameters to tune and does not do distributional forecasting. As a result, the validation MPIW and PICP metrics are omitted as well as the selected hyperparameters. Still, the median site-wise RMSEs at 1 and 12 week horizons were 0.026 and 0.104, respectively.

4.1.2. N-HiTS

The selected hyperparameters for the N-HiTS model after 30 trials with Optuna minimizing the quantile loss for a horizon 12 forecast are in Table 4.1. The median RMSEs were 0.022

Table 4.1.: Hyperparameter Selection for N-HiTS

Hyperparameter	Value
Number of Steps	3000
Learning Rate (η)	7.07×10^{-5}
Batch Size	128

(h1) and 0.060 (h12), the median MPIWs were 0.114 (h1) and 0.301 (h12), and the median PICPs were 0.977 (h1) and 0.976 (h12).

4.1.3. Chronos2 Zero-Shot

Because of the zero-shot capability of this model, there are no optimal hyperparameters to report. This method is, however, capable of doing distributional forecasting and the median validation set metrics are as follows: The median RMSEs were 0.030 (h1) and 0.096 (h12), the median MPIWs were 0.361 (h1) and 0.643 (h12), and the median PICPs were 1.000 (h1) and 0.997 (h12).

4.1.4. Chronos2 Finetuned

The hyperparameters of Chronos2 with finetuning after 30 trials of hyperparameter optimization are given in Table 4.2. The validation set metrics associated with these

Table 4.2.: Hyperparameter Selection for Chronos2

Hyperparameter	Value
Learning Rate (η)	5×10^{-6}
Number of Steps	3000
Batch Size	100
LoRA	False
LoRA α	N/A
LoRA r	N/A

hyperparameters are as follows: The median RMSEs were 0.023 (h1) and 0.062 (h12), the median MPIWs were 0.135 (h1) and 0.336 (h12), and the median PICPs were 0.988 (h1) and 0.989 (h12).

4.1.5. Combined Results

The following section aggregates the results of the temporal modeling on the Test-Train set in two figures with metrics computed site-wise. Figure 4.1 depicts the metrics for a forecast horizon of 1 week and Figure 4.2 depicts the metrics for a forecast horizon of 12 weeks. Tables with additional results on the temporal test set are detailed in Appendix D.1.

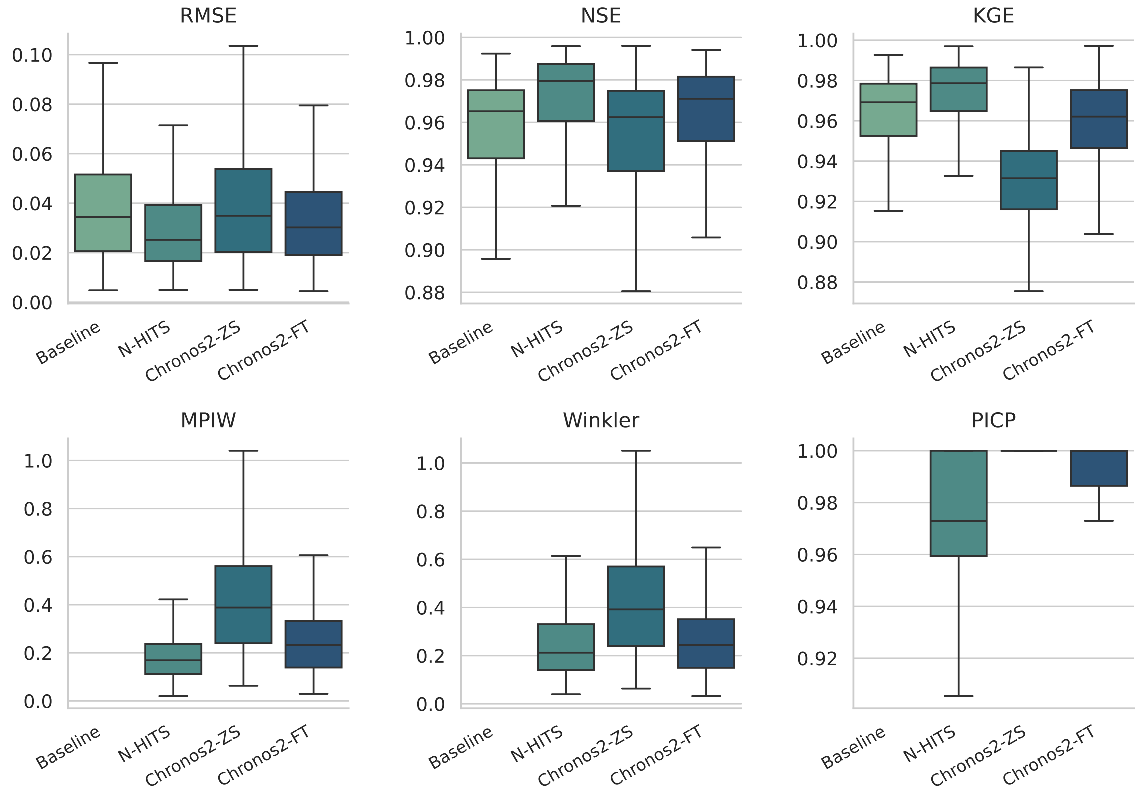

Figure 4.1.: Temporal Model Test-Train Horizon 1 Boxplots. N-HiTS outperforms all other models in all metrics with PICP closer to the specified confidence level. The baseline is strong at this horizon though without distributional forecasting ability.

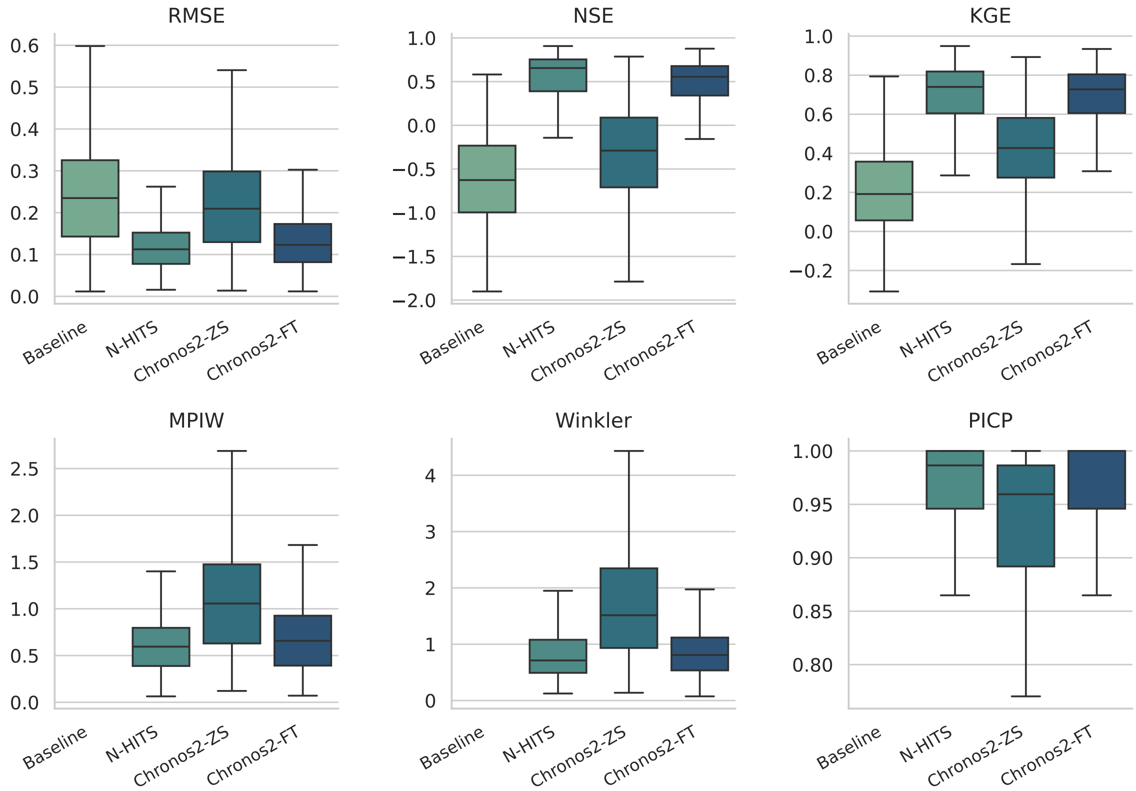

Figure 4.2.: Temporal Model Test-Train Horizon 12 Boxplots. N-HiTS outperforms all other models in all metrics. There is a larger gap in performance between the baseline and the other methods. Additionally, there is a reduction in performance relative to the horizon 1 forecasts.

4.2. Spatial

The results for the spatial models are computed using the Train-Train data and then evaluated on the Val-Val and Test-Test data. This ensures that the spatial training set is utilized for training the models and the spatial validation and spatial test sets are used for model selection and evaluation, respectively. Additionally, since the temporal splits are chosen to be validation and train, comparisons to the spatiotemporal models in further sections can be made easily. The number of sites in the Val-Val and Test-Test are the same as those in the Train-Val and Train-Test so the accuracy of metric values is not sacrificed by using a smaller sample in their computation. Finally, utilizing spatial models on future timesteps, effectively neglecting the micro fluctuations of the target with respect to time, can serve as a long-term forecast.

The section proceeds as follows, it presents any hyperparameters relevant to the spatial model and the validation metrics associated with selecting these. Thereafter it presents a combined analysis of the spatial models on the Test-Test data. It is noteworthy that since we are focused primarily on the spatial generalization of the models, the model predictions will be compared to the mean groundwater level over time at the evaluation sites. That is, we "average-out" the time dimension from the analyses so the spatial models in this section are not evaluated in the same manner as the spatiotemporal models. This allows us to better understand how well these models are able to capture the "offset" at each unseen location. Additionally, this setup more closely resembles what a scientist may do if they are interested in determining the groundwater level at a given site and look up the values on a website using something such as groundwater contour lines. Correspondingly, residuals are of the form: $e_s = \bar{y}_s - \hat{y}_s$ where s denotes the site. Additionally, the intervals for the Gaussian Processes used in the calculation of uncertainty metrics are computed: $L = \hat{y} - z_{1-\alpha/2} \cdot \hat{\sigma}$ and $U = \hat{y} + z_{1-\alpha/2} \cdot \hat{\sigma}$, the ResNet is trained to output quantile predictions outlined in Paragraph 3.2.2. The use of the same confidence level as predicted quantiles increases comparability.

4.2.1. K-Nearest Neighbors

Using the elbow method, the optimal K is 2 from a search space of K ranging from 1 to 10. The metrics on the val-val set with this hyperparameter are: 3.356 (spatial MAE), 6.994 (spatial RMSE), and 0.897 (spatial NSE). There are no probabalistic evaluation metrics since the method only does point prediction. This is the lowest validation set performance across all metrics for all of the spatial models and serves as a spatial baseline method.

4.2.2. TabPFN

TabPFN is a foundation model that does in context learning and therefore has no hyperparameters to tune. Still, the metrics on the validation set are: 1.816 (spatial MAE), 2.968 (spatial RMSE), and 0.981 (spatial NSE) with no probabalistic evaluation metrics. TabPFN has the best validation set spatial MAE of all models tested. There were never plans to build a full spatiotemporal model with TabPFN since it isn't very flexible in terms of software publicity which makes modification of its internals challenging.

4.2.3. Fourier Feature Expanded Residual Network

The optimal hyperparameters for the Fourier Feature Expansion Spatial ResNet after 30 trials of hyperparameter optimization are given in Table 4.3.

Table 4.3.: FFES-ResNet Optimized Hyperparameters

Hyperparameter	Optimized Value
Fourier Dimension	128
Hidden Dimension	455
Number of Blocks	4
Fourier Scale (f_{scale})	0.135
Learning Rate (lr)	8.108×10^{-4}

The validation metrics are: 1.875 (spatial MAE), 2.805 (spatial RMSE), 0.983 (spatial NSE), 33.871 (PICP 98%), 1.591 (MPIW), and 124.295 (Winkler) This model has the highest validation set performance in all metrics aside from spatial MAE where TabPFN outperforms it.

4.2.4. Gaussian Process on Coordinates

The hyperparameters for the Gaussian process on coordinates are given in Table 4.4. The optimization of Gaussian process hyperparameters happens internally in the Gpytorch library and is specified in Section 3.2.2.

Table 4.4.: Hyperparameters for Gaussian Process on Coordinates

Hyperparameter	Optimized Value
Spatial Lengthscale ℓ_x	0.2419
Spatial Lengthscale ℓ_y	0.2801
Noise Variance σ_n^2	0.0169

The validation metrics are: 2.558 (spatial MAE), 4.780 (spatial RMSE), 0.952 (spatial NSE), 100.000 (PCIP 98%), 26.280 (MPIW), 26.280 (Winkler). The spatial lengthscales are not equal, indicating anisotropy.

4.2.5. Gaussian Process on Coordinates and Geologic Principal Components

The hyperparameters for the Gaussian process on coordinates and geologic principal components are given in Table 4.5.

The validation metrics are: 2.237 (spatial MAE), 3.370 (spatial RMSE), 0.976 (spatial NSE), 93.548 (PCIP 98%), 10.952 (MPIW), 24.332 (Winkler) The spatial lengthscales are larger than those of the Gaussian process on coordinates only. Additionally, the relative ordering of the coordinate lengthscales has switched while remaining unequal, indicating

Table 4.5.: Hyperparameters for Gaussian Process on Coordinates and Geologic Principal Components

Hyperparameter	Optimized Value
Spatial Lengthscale ℓ_x	0.5497
Spatial Lengthscale ℓ_y	0.5066
Environmental Lengthscale ℓ_{PC1}	2.6061
Environmental Lengthscale ℓ_{PC2}	4.6778
Environmental Lengthscale ℓ_{PC3}	5.2399
Environmental Lengthscale ℓ_{PC4}	14.9591
Environmental Lengthscale ℓ_{PC5}	2.3071
Environmental Lengthscale ℓ_{PC6}	9.2479
Environmental Lengthscale ℓ_{PC7}	13.6081
Environmental Lengthscale ℓ_{PC8}	1.9709
Noise Variance σ_n^2	0.0012

anisotropy. The smallest lengthscale corresponding to the principal components is the one associated with principal component 8. Additionally the noise variance is lower than that of the Gaussian process on coordinates only.

4.2.6. Gaussian Process on Coordinates and Reduced Tessera Embeddings

The hyperparameters of the Gaussian process on coordinates and reduced Tessera embeddings are given in Table 4.6.

Table 4.6.: Hyperparameters for Gaussian Process on Coordinates and Reduced Tessera Embeddings

Hyperparameter	Optimized Value
Spatial Structure (k_0) Lengthscale ℓ_x	0.2382
Spatial Structure (k_0) Lengthscale ℓ_y	0.2785
GeoTessera (k_1) Lengthscales $\ell_{0\ldots19}$ (Range)	$\sim 2.8524 - 3.1548$
Noise Variance σ_n^2	0.0124

The validation metrics are: 2.691 (spatial MAE), 4.682 (spatial RMSE), 0.954 (spatial NSE), 100.000 (PCIP 98%), 26.029 (MPIW), 26.029 (Winkler) The lengthscales are smaller than the Gaussian process with geologic principal components indicating that there is less signal in the reduced tessera embeddings alone. Additionally, the lengthscales of the 20 reduced tessera embeddings have lengthscales approximately 10 times as large as the coordinate length scales, indicating the model is primarily reliant on the coordinates.

4.2.7. Gaussian Process on Coordinates, Geologic Principal Components, and Reduced Tessera Embeddings

The hyperparameters of the Gaussian process on coordinates, geologic principal components, and reduced Tessera embeddings are given in Table 4.7.

Table 4.7.: Hyperparameters for Gaussian Process on Coordinates, Geologic Principal Components, and Reduced Tessera Embeddings

Hyperparameter	Optimized Value
Spatial Structure (k_0) Lengthscale ℓ_x	0.3613
Spatial Structure (k_0) Lengthscale ℓ_y	0.3635
Environmental (k_1) Lengthscales (Range)	$\sim 2.5719 - 18.4801$
GeoTessera (k_2) Lengthscales $\ell_{0\ldots19}$ (Range)	$\sim 3.4362 - 3.7897$
Noise Variance σ_n^2	0.0013

The validation metrics are: 2.107 (spatial MAE), 3.115 (spatial RMSE), 0.980 (spatial NSE), 95.161 (PCIP 98%), 11.016 (MPIW), 19.408 (Winkler) This is the best performing model in terms of spatial MAE, spatial RMSE, spatial NSE, and Winkler score on the validation set. The spatial lengthscales are equal indicating isotropy.

4.2.8. Combined Results

The results on the Test-Test set of each model provided the Train-Train set are given in Table 4.8. In the test set metrics, the ordering of the Gaussian processes performance reverses and the ResNet and TabPFN are worse than all Gaussian processes. The K-Nearest Neighbors method is still the worst performing.

Table 4.8.: Spatial Generalization Metrics Across Unseen Test Site Means

Model	Spatial MAE	Spatial RMSE	Spatial-NSE	PICP (%)	MPIW	Winkler Score
gpr-coords	2.173	3.603	0.979	95.161	24.992	34.471
gpr-tessera	2.332	3.755	0.978	95.161	24.898	36.826
gpr-pcs	2.483	5.020	0.960	91.936	12.572	59.511
gpr-tessera-pcs	2.562	5.322	0.955	91.936	12.631	64.857
ffe-res-net	2.550	5.354	0.954	35.484	1.502	196.785
tab-pfn	2.457	5.888	0.945	—	—	—
knn	3.337	6.003	0.943	—	—	—

4.3. Decoupled Spatiotemporal

Models in this section consist of forecasters which make forecasts on the Train-Val and Train-Test sets with a Gaussian process fit on the forecasted values at each horizon and timestep to predict the groundwater level in the Val-Val and Test-Test sets. The main point of exploration in this section is to better understand how the different kernels and input data affect the forecast performance at unknown sites. This section is also the first for which a full, spatiotemporal model is evaluated. That is, the first section where the forecast performance at sites without histories is evaluated.

For each of the models, the section proceeds as follows, the hyperparameters on the validation set are presented, these directly affected model choice and design. It is noteworthy that due to the large number of Gaussian processes fit, distributions of hyperparameters for each set of input features are given. The tables of validation metrics computed on the Val-Val set are provided for each of the metrics, forecasters, and horizons in the following Gaussian process variants. Boxplots of the performance on the Test-Test set are given for each of the temporal models, each of the horizons, and each of the metrics Tables corresponding to the figures shown in the following subsections are given in Appendix D.2.

4.3.1. Gaussian Process on Coordinates

This subsection covers the results for spatiotemporal forecasting when only the geographic coordinates are provided to the Gaussian processes. In pursuit of brevity, we omit the lengthscale analysis of these models and refer the reader to Subsection 4.2.4 for an analysis of coordinate-only lengthscales. We provide the distribution of the noise variances for the Gaussian processes on coordinates over all Val-Val timesteps in Table 4.9. The table shows

Table 4.9.: Noise Variance Distributions for the Coordinate-Only Model on Val-Val

Hyperparameters	Mean	Std	Median	Q25	Q75
Noise Variance σ_n^2 (H1)	0.0172	0.0001	0.0171	0.0171	0.0173
Noise Variance σ_n^2 (H12)	0.0171	0.0001	0.0172	0.0170	0.0172

little change between the noise variances of the Gaussian processes fit on horizon 1 and horizon 12 forecasts. In Table 4.10 we see the Val-Val metrics for each of the decoupled models with a Gaussian process given only coordinates.

In Figure 4.3 the performance metrics on the Test-Test set with only coordinates as input are given for horizons 1 and 12 weeks. The boxplots show stability in all metrics except KGE between the horizon 1 and horizon 12 spatiotemporal forecasts with little difference at the median. The KGE displays variation between the horizon 1 and horizon 12 forecasts with a ranking of the underlying temporal models similar to that of the previous temporal section at horizon 12.

Table 4.10.: Val-Val Metrics for Coordinate-Only Gaussian Process on Forecasters

Model	Metric Horizon	RMSE Median	RMSE IQR	NSE Median	NSE IQR	KGE Median	KGE IQR
Baseline	H1	1.335	2.265	-156.106	935.899	0.497	0.698
	H12	1.339	2.280	-108.556	861.280	0.008	0.562
N-HiTS	H1	1.337	2.266	-149.976	889.208	0.488	0.651
	H12	1.328	2.283	-100.676	860.598	0.420	0.501
C2 Zero Shot	H1	1.319	2.265	-153.859	943.913	0.511	0.642
	H12	1.336	2.284	-105.739	872.005	0.148	0.619
C2 Finetuned	H1	1.335	2.267	-161.606	926.759	0.495	0.661
	H12	1.321	2.298	-101.257	902.402	0.410	0.494

Model	Metric Horizon	MPIW Median	MPIW IQR	PICP Median	PICP IQR	Winkler Median	Winkler IQR
Baseline	H1	21.373	10.931	1.000	0.000	21.373	10.931
	H12	21.373	10.931	1.000	0.000	21.373	10.931
N-HiTS	H1	21.285	10.746	1.000	0.000	21.285	10.746
	H12	21.254	10.697	1.000	0.000	21.254	10.697
C2 Zero Shot	H1	21.404	10.981	1.000	0.000	21.404	10.981
	H12	21.395	10.943	1.000	0.000	21.395	10.943
C2 Finetuned	H1	21.231	10.699	1.000	0.000	21.231	10.699
	H12	21.360	10.945	1.000	0.000	21.360	10.945

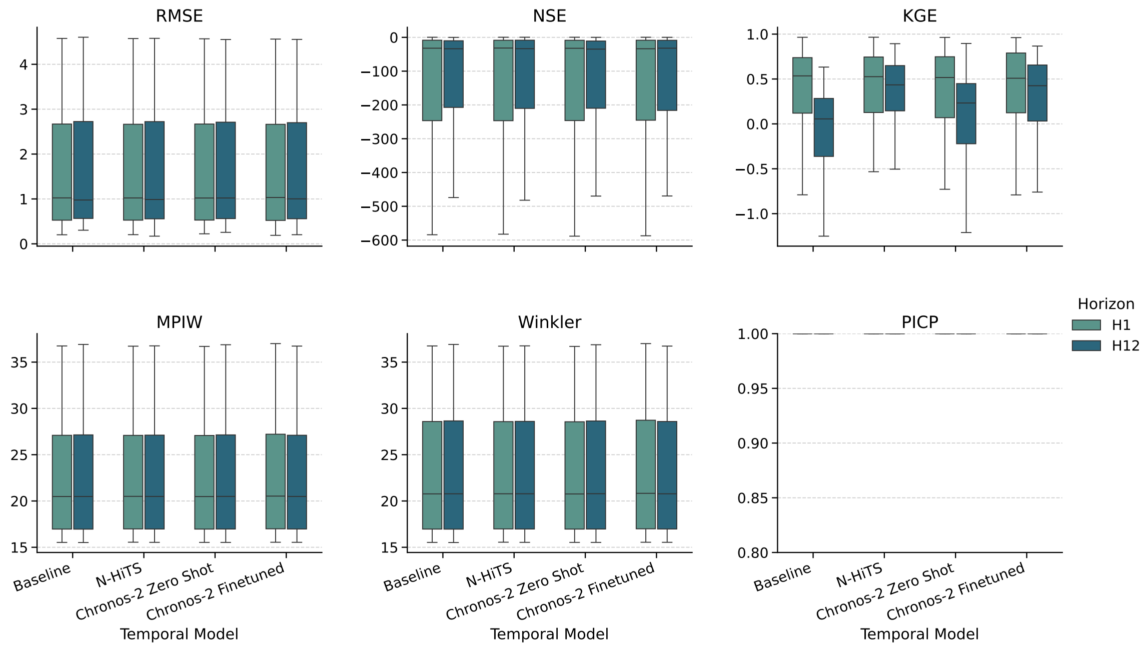

Figure 4.3.: GP on Coordinates Test-Test Performance Metrics Boxplots. The models have similar perfromances regardless of the horizon or the forecaster chosen (except in the case of KGE). There is full coverage likely due to large predictive intervals.

4.3.2. Gaussian Process on Coordinates and Geologic Principal Components

This section presents results of the spatiotemporal forecasting when both geographic coordinates and geologic principal components are provided via two additive kernels to the Gaussian processes. The distributions of the noise variances for each of the Gaussian processes fit in this data setting are given in Table 4.11. As in the section previous, the

Table 4.11.: Noise Variance Distributions for the Coordinate and Geologic Principal Components Model on Val-Val

Hyperparameters	Mean	Std	Median	Q25	Q75
Noise Variance σ_n^2 (H1)	0.0010	0.0001	0.0009	0.0009	0.0010
Noise Variance σ_n^2 (H12)	0.0010	0.0001	0.0010	0.0009	0.0010

noise variance stays similar between the horizon 1 and horizon 12 forecasts.

The distributions of lengthscales for each of the Gaussian processes fit on coordinates and geologic principal components are displayed in the boxplots of Figure 4.4. Due to the structure of lengthscales in Gaussian process kernels, lower values indicate higher importance of features. The geographic coordinates have the lowest values with the least spread of any other feature, indicating they are the most important in the spatial interpolation portion of the spatiotemporal forecasting. Of the geologic principal components, principal components 8, 1, and 5 have the lowest values indicating they are the most important principal components in the spatial interpolation step of making spatiotemporal forecasts. Further analysis of success and failure modes of the models using the principal components are examined in further sections.

Table 4.12 shows the metrics on the Val-Val set for the models with geographic coordinates and geologic principal components.

In Figure 4.5 the results on the Test-Test set with coordinates and geologic principal components as input to the Gaussian processes are given for horizons 1 and 12 weeks. The boxplots exhibit similar behavior to those discussed in the previous subsection. Notably, the performances are slightly worse than the model given only coordinates except for the width of prediction intervals where the model maintains 100% coverage but has smaller prediction intervals, measured with MPIW.

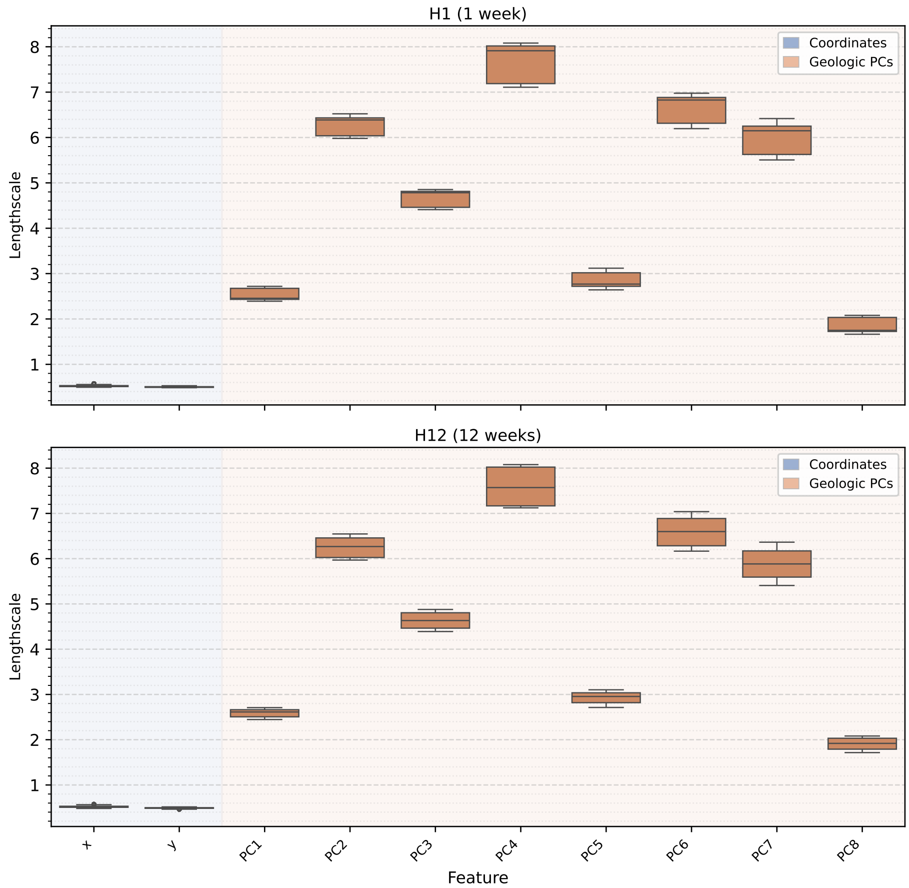

Figure 4.4.: GP on Coordinates and Geo PCs Val-Val Lengthscale Boxplots. The most important features based on the lengthscale analysis are the geographic coordinates with the 8th, 1st, and 5th principal components showing large importances. Principal component 4 is the least relevant based on this analysis.

Table 4.12.: Val-Val Metrics for Coordinate and Geologic Principal Component Gaussian Process on Forecasters

	Metric	RMSE		NSE		KGE	
Model	Horizon	Median	IQR	Median	IQR	Median	IQR
Baseline	H1	1.405	2.436	-118.254	764.469	0.440	0.758
	H12	1.421	2.435	-98.724	722.243	-0.046	0.527
N-HiTS	H1	1.216	2.291	-82.060	725.025	0.455	0.659
	H12	1.199	2.323	-66.446	697.517	0.292	0.721
C2 Zero Shot	H1	1.416	2.442	-121.283	771.808	0.343	0.856
	H12	1.297	2.296	-84.121	696.330	0.054	0.701
C2 Finetuned	H1	1.401	2.439	-120.076	766.622	0.443	0.761
	H12	1.410	2.375	-83.642	715.072	0.348	0.738

	Metric	MPIW		PICP		Winkler	
Model	Horizon	Median	IQR	Median	IQR	Median	IQR
Baseline	H1	10.292	7.489	1.000	0.000	10.600	8.251
	H12	10.292	7.489	1.000	0.000	10.600	8.251
N-HiTS	H1	10.259	7.172	1.000	0.000	11.043	8.260
	H12	10.189	7.166	1.000	0.000	10.987	8.151
C2 Zero Shot	H1	10.444	7.341	1.000	0.000	10.638	8.134
	H12	10.107	7.281	1.000	0.000	10.734	8.561
C2 Finetuned	H1	10.304	7.423	1.000	0.000	10.566	8.243
	H12	10.265	7.679	1.000	0.000	10.624	8.441

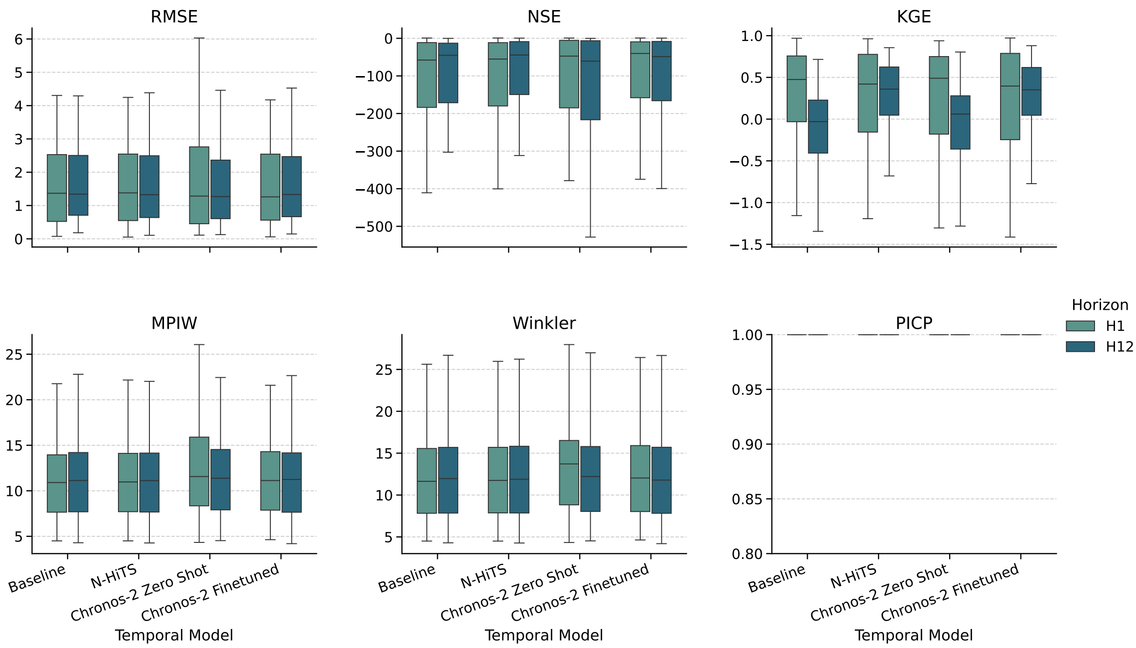

Figure 4.5.: GP on Coordinates and Geo PCs Test-Test Performance Metrics Boxplots. There is variation between the forecast models with the N-HiTS H12 slightly outperforming the others in all metrics except RMSE and PICP where they are approximately equal. The PICP is still 100% likely due to the large MPIW though the MPIW is approximatley half of that when no PCs are provided to the GP.

4.3.3. Gaussian Process on Coordinates, Geologic Principal Components, and Reduced Tessera Embeddings

This section presents results of the spatiotemporal forecasting when geographic coordinates, geologic principal components, and reduced Tessera embeddings are provided via three additive kernels to the Gaussian processes. The distributions of the noise variances for each of the Gaussian processes fit in this data setting are given in Figure 4.13. The distributions of the noise variances here exhibit larger differences than in the previous subsections. Specifically, the median of the horizon 12 noise variance distribution is smaller than that of the horizon 1 noise variance distribution.

Table 4.13.: Noise Variance Distributions for the Coordinate, Geologic Principal Components, and Reduced Tessera Embeddings Model on Val-Val

Hyperparameters	Mean	Std	Median	Q25	Q75
Noise Variance σ_n^2 (H1)	0.0020	0.0007	0.0019	0.0013	0.0027
Noise Variance σ_n^2 (H12)	0.0010	0.0001	0.0010	0.0009	0.0010

The distributions of lengthscales for each of the Gaussian processes fit on coordinates, geologic principal components, and GeoTessera principal components are displayed in the boxplots of Figure 4.6. Due to the large number and similar behavior of the Tessera principal components (20), the lengthscales of each are aggregated into one boxplot. Summary statistics of the distributions of each are provided in tables in Appendix Subsection D.2.3. The distributions exhibit less variance of lengthscales between features in this setting. In both the horizon 1 and 12, the Tessera principal components have a lower whisker slightly larger than the value 4, with 5 other features displaying higher importance than the majority of the most important Tessera features. Additionally, geologic principal component 1 has a smaller median than principal component 5 and 8 at horizons 1 and 12 indicating higher importance which is a different ordering than the lengthscales without Tessera features. Still, the geographic coordinates are the most important with y attaining smaller values than x in both horizons.

In Table 4.14, the metrics on the Val-Val set are reported.

In Figure 4.7 the results on the Test-Test set with coordinates and geologic principal components as input are given for horizons 1 and 12 weeks. The performances differ slightly from the aforementioned settings with median RMSE across horizons being larger in this setting. The median KGE across horizons is larger than the two previously mentioned settings and, while maintaining 100% coverage, the MPIW is slightly smaller than the Gaussian process on coordinates though higher than the Gaussian process on coordinates and geologic principal components.

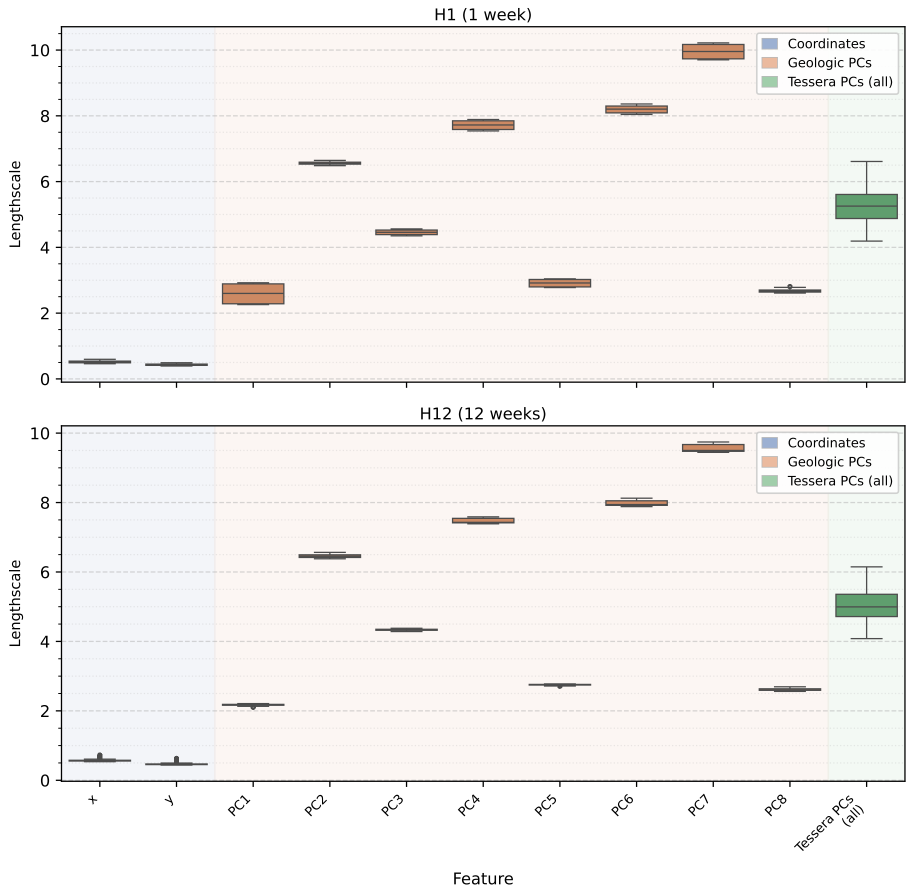

Figure 4.6.: GP on Coordinates, Geo PCs, and Tessera PCs Val-Val Lengthscale Boxplots. The most important features based on the lengthscales are the geographic coordinates. They are followed by PCs 1, 8, and 5. The Tessera PCs have importance situated in the center of the geologic PCs with 4 more important and 4 less important.

Table 4.14.: Val-Val Metrics for Coordinate, Geologic Principal Component, and Reduced Tessera Embeddings Gaussian Process on Forecasters

Model	Metric Horizon	RMSE Median	RMSE IQR	NSE Median	NSE IQR	KGE Median	KGE IQR
Baseline	H1	1.399	2.756	-178.379	805.128	0.463	0.648
	H12	1.677	2.679	-142.836	777.936	0.009	0.587
N-HiTS	H1	1.415	2.756	-179.239	767.899	0.547	0.714
	H12	1.499	2.479	-125.102	710.613	0.352	0.653
C2 Zero Shot	H1	1.286	2.564	-162.341	809.100	0.486	0.808
	H12	1.548	2.520	-136.059	714.264	-0.041	0.744
C2 Finetuned	H1	1.482	2.573	-162.199	915.076	0.461	0.687
	H12	1.494	2.522	-132.493	734.898	0.371	0.609

Model	Metric Horizon	MPIW Median	MPIW IQR	PICP Median	PICP IQR	Winkler Median	Winkler IQR
Baseline	H1	17.297	5.838	1.000	0.000	17.599	6.141
	H12	17.761	6.128	1.000	0.000	18.035	6.638
N-HiTS	H1	17.341	5.824	1.000	0.000	17.640	6.094
	H12	18.051	6.584	1.000	0.000	18.357	7.043
C2 Zero Shot	H1	17.274	6.116	1.000	0.000	17.864	6.554
	H12	18.008	6.624	1.000	0.000	18.308	7.058
C2 Finetuned	H1	17.802	6.505	1.000	0.000	18.172	6.919
	H12	18.103	6.608	1.000	0.000	18.392	7.074

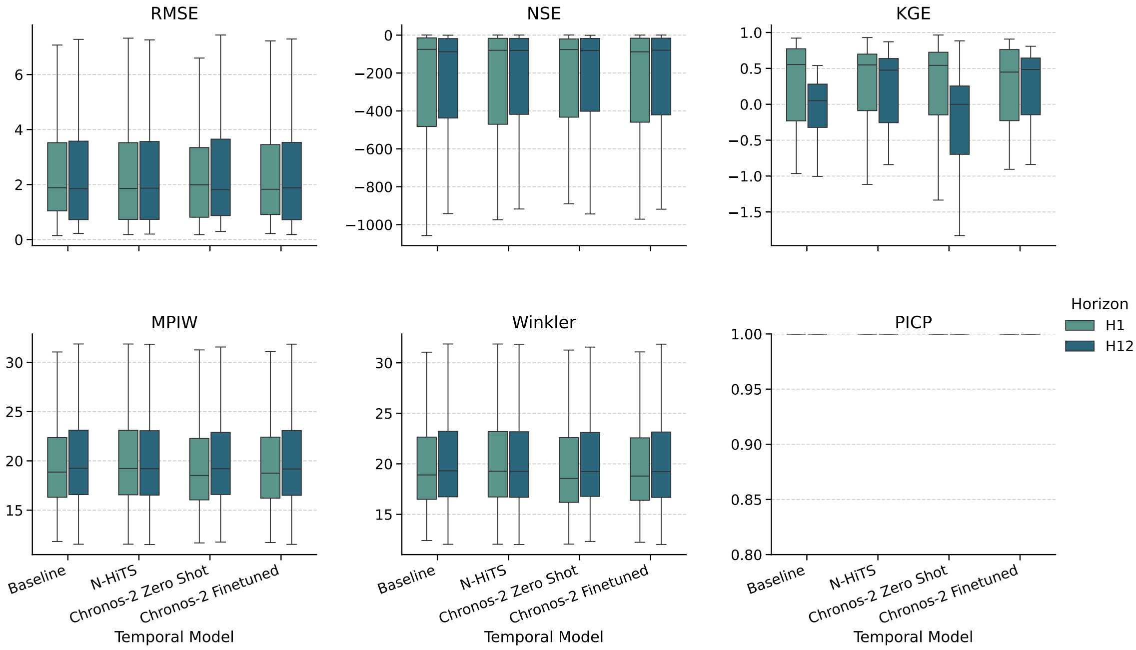

Figure 4.7.: GP on Coordinates, Geo PCs, and Tessera PCs Test-Test Performance Metrics Boxplots. The performance collapses to approximate equality throughout the horizons, methods, and metrics except in the case of KGE where N-HiTS is still the top perfromer and the quality degrades as the horizon increases. There is also a slight increase in Winkler as the horizon increases.

4.4. Coupled Spatiotemporal

This section examines methods where the spatial and temporal modeling is combined within the models themselves. It differs from decoupled modeling by using one method to jointly learn spatial and temporal patterns. Again, the models herein seek to produce spatiotemporal forecasts, that is, forecasts at sites with no historical values of the target variable. The models considered consist of the Kronecker Gaussian Process, the Cross Attention Model for the fusion of forecasts, and the Cross Attention Model with the Variational Gaussian Process.

The section proceeds similarly to previous sections with the hyperparameters from the validation set, then model performances on the test sets are examined with boxplots. Additional tables are provided in the Appendix Section D.3 which give numbers used to create the performance metric boxplots. Finally, since there are far fewer models fit, there is no need to use distributions to examine the hyperparameters of the models examined.

4.4.1. Kronecker Gaussian Process

As mentioned, this model does not use a forecast horizon and instead acts as a functional map from a given 3-tuple of (x,y,t) to the predicted groundwater level. Therefore, there is no analysis done on the forecast horizon. Additionally, analysis of the Tessera embeddings was omitted due to time constraints and the marginal performance gains in other models.

The hyperparameters of the Kronecker Gaussian Process on Coordinates and Time on Test-Test are given in Table 4.15. The table shows a much larger lengthscale for the

Table 4.15.: Lengthscale and Noise Variance Hyperparameters of Kronecker Gaussian Process on Coordinates and Time

Spatial Lengthscale ℓ_x	Spatial Lengthscale ℓ_y	Temporal Lengthscale ℓ_t	Noise Variance σ_n^2
0.178	0.139	30.177	0.0103

temporal variable, indicating it has less importance to the model than the spatial variables.

The hyperparameters of the Kronecker Gaussian process on coordinates, time, and geologic principal components are given in Table 4.16. The table shows a similar ranking

Table 4.16.: Lengthscale and Noise Variance Hyperparameters of Kronecker Gaussian Process on Coordinates, Time, and Geologic Principal Components on Val-Val

ℓ_x	ℓ_y	ℓ_l	ℓ_{PC_1}	ℓ_{PC_2}	ℓ_{PC_3}	ℓ_{PC_4}	ℓ_{PC_5}	ℓ_{PC_6}	ℓ_{PC_7}	ℓ_{PC_8}	σ_n^2
0.283	0.290	37.685	2.570	6.775	3.355	8.001	3.442	7.932	8.071	2.013	0.0103

as the previous sections for the lengthscales with the coordinates being most important followed by PCs 8, 1, and 5. Again the temporal lengthscale is very large indicating less importance in the model

In Table 4.17, we see the metrics on the validation set. One can see the RMSE remains

Table 4.17.: Kronecker GP Val-Val Performance Metrics — Median [IQR]

Model	RMSE	NSE	KGE	MPIW	Winkler	PICP
KGP No-PC	1.141 [2.779]	-85.200 [587.631]	0.039 [0.299]	32.364 [57.131]	33.868 [55.665]	1.000 [0.000]
KGP PC	1.556 [2.621]	-111.054 [740.738]	0.043 [0.380]	9.650 [8.225]	11.979 [10.058]	1.000 [0.000]

low with a lower value in the Kronecker Gaussian process without PCs than the one with them. Contrastively, the MPIW is lower in the Kronecker Gaussian Process with geologic PCs while maintaining high PICP coverage, yielding a lower Winkler score as well.

The results on the test set are given in Figure 4.8. Between the models fit with and

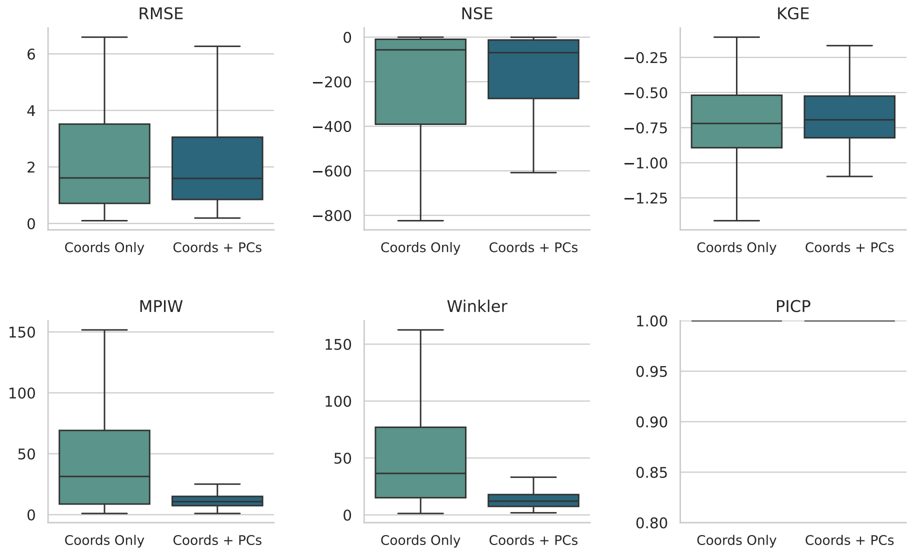

Figure 4.8.: Kronecker Gaussian Process on Test-Test Data. Adding the geologic PCs increases the predictive performance in all metrics with those evaluating the distributional forecasts seeing the most improvement. There is still 100% coverage presumably due to the large size of the prediction intervals shown in MPIW.

without geologic principal components, the median RMSE appears to be similar. When principal components are introduced, the variances in all of the metric distributions are smaller and the predicitive uncertainty, quantified by the MPIW is smaller while maintaining 100% coverage. Finally, additional metrics are provided in Appendix D.3.1.

4.4.2. Attention

In this section we examine the results of the models utilizing attention. First the Cross Attention model with quantile output is examined, then the model with the Variational Gaussian process output. Combined results are given at the end.

Hyperparameters The optimized hyperparameters of the quantile output Cross Attention model are listed in Table 4.18. It is noteworthy that the dropout rate is approximately 30%

Table 4.18.: Cross Attention Selected Hyperparameters

Hyperparameter	Value
Embedding Dimension	256
Number of Heads	2
Attention Layers	2
Learning Rate (lr)	0.001701
Dropout	0.301791

which could indicate overfitting.

The optimal hyperparameters for the Gaussian Process with Cross Attention are listed in Table 4.19. It is again noteworthy that the dropout rate is approximately 39%, close to

Table 4.19.: Gaussian Process with Cross Attention Selected Hyperparameters

Hyperparameter	Value
Embedding Dimension	128
Number of Heads	4
Attention Layers	2
Learning Rate (lr)	0.008132
Dropout	0.392565
Inducing Points (n_{inducing})	128

the upper limit of 40%, which could indicate overfitting.

Performance Metrics The performance metrics of the Attention models on the Validation set are given in Table 4.20. From the table one can see low RMSE in both models across the

Table 4.20.: Attention Metrics for Val-Val (Horizons 1 and 12) — Median [IQR]

Horizon	Model	RMSE	NSE	KGE	MPIW	Winkler	PICP
1	Attention	1.195 [1.630]	-89.395 [902.912]	-0.585 [0.242]	1.328 [0.681]	44.337 [139.457]	0.000 [0.857]
	GP-Attention	1.146 [1.588]	-94.898 [985.793]	-0.414 [0.001]	1.355 [0.000]	47.585 [149.908]	0.000 [0.767]
12	Attention	1.135 [1.611]	-74.742 [693.091]	-0.463 [0.191]	1.402 [0.679]	40.111 [146.596]	0.000 [0.807]
	GP-Attention	1.151 [1.794]	-84.409 [694.740]	-0.415 [0.003]	1.420 [0.000]	43.856 [161.005]	0.000 [0.833]

horizons is obtained though the KGE is negative indicating poor forecasting performance. Additionally, the PICP has a median of 0 presumably due to the small values of the MPIW metric across the models and horizons.

The performance metrics of the Cross Attention and Gaussian Process with Cross Attention models on the Test set are given in Figure 4.9. Between horizons, there are larger differences present in the metrics of the attention model than those of the GP-Attention model. In terms of RMSE, the attention model slightly outperforms the Gaussian process attention model. The KGE of the GP-Attention model has very little spread about -0.41

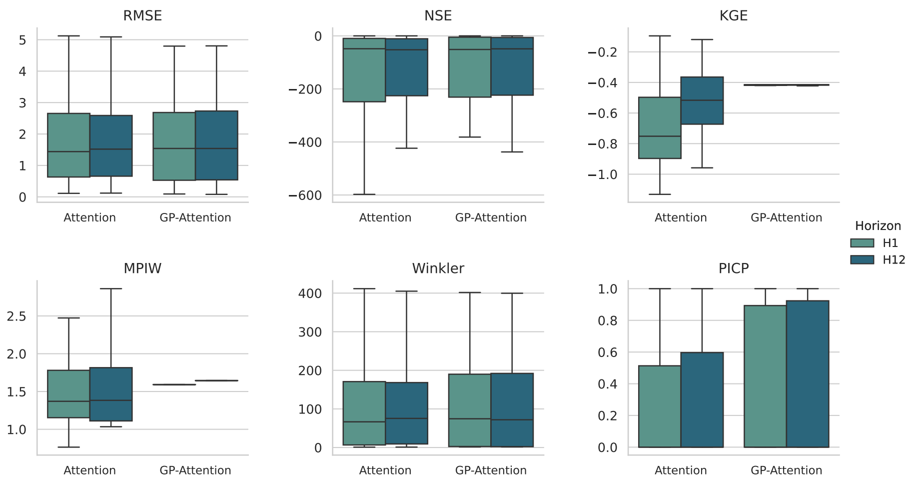

Figure 4.9.: Cross Attention and Gaussian Process with Cross Attention on Test-Test Data. The performance between horizons is approximately the same for both models in RMSE and NSE. The KGE sees an improvement in the attention only model as the horizon increases while the GP with Attention stays the same with little spread. The PICP has a median of 0 presumably due to the very small values of MPIW.

and outperforms the attention model at both horizons. The MPIW of both models is very small relative to the previous models examined. The small intervals coupled with mediocre point estimate metrics manifests itself as little coverage and consequently large values of the Winkler Score and small values of the PICP.

4.5. Best Spatiotemporal Models & Relative Comparisons

We examine the Gaussian Process with Principal Components on N-HiTS forecasts due to the best RMSE performance on the validation set with medians of 1.216 (H1) and 1.199 (H12). While this is not the overall best RMSE on the Val-Val set, it maintained high KGE with only the Chronos-2 Finetuned model besting it in the horizon 12 forecasts. Moreover, the PICP and Winkler scores were better than the other examined models. Finally, the Gaussian process based models are popular in hydrological machine learning and more analysis could be of interest to the community.

The effect of the principal components on model performance is evaluated by comparing the models with to those without. Visualizing the model predictions ordered by the corresponding metric scores is done to gain a better visual understanding of the numbers for NSE and RMSE. Finally, examining how the performance of the spatial model changes when fit on true values and historical means instead of forecasts is explored.

4.5.1. Perfromances on Geographic Maps

The performance of the spatiotemporal forecasts with respect to the spatial distribution of wells is examined by means of geographic maps. In the following, maps using the same coordinate reference system as those in Section 3.1 are used to examine the effect of both spatial data density and location on the models. The first row consists of GPs fit without the use of geologic principal components and the second row consists of GPs fit with them. The first column shows the performances of GPs fit on true groundwater levels at locations in the spatial training set. The second column shows the performances of GPs fit on N-HITS forecasts with horizon 1. The final column shows the performances of GPs fit on N-HITS forecasts with horizon 12. Each map has a blue circle denoting the best performing site and a red circle denoting the worst performing site with respect to the metric in the plots. Finally, each figure shows maps with the performance metrics RMSE and Winkler, to assess the deviation from the true values and the model's distributional forecast performance. These metrics were chosen in pursuit of terseness and maximal interpretability.

In Figure 4.10, there is agreement between the worst performing sites when either including or excluding geologic principal components. The best performing sites tend to be in a higher spatial data density area though in the bottom rightmost plot this is not the case. The distance between the best and worst performing sites is large relative to the distances to the nearest sites and there appears to be some clustering of poor performances. There is improvement in the performance when geologic principal components are added in several locations in the south of Brandenburg though there is also degredation in several northern locations.

In Figure 4.11, there are larger outliers overall. In the case of the GP with geologic principal components, there is agreement between the best and worst performing sites. In the case of the GP without geologic principal components, the best and worst sites are relatively close to eachother given the scale of the data. When principal components are introduced, the geographic distance between the best and worst sites is larger in the three scenarios though there are larger outliers in terms of the overall Winkler score.

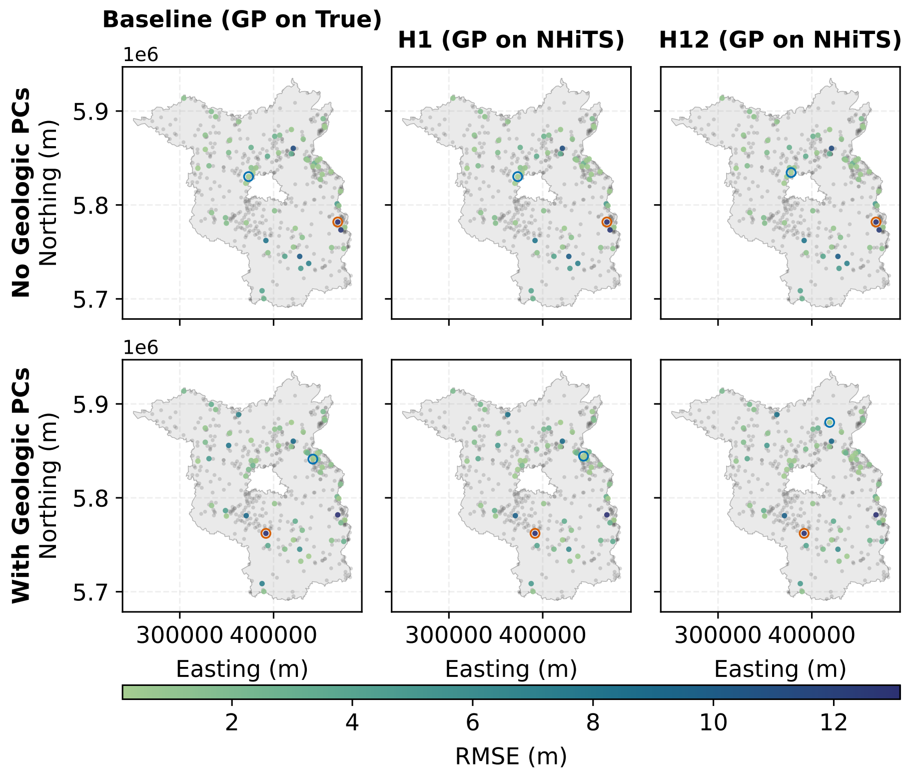

Figure 4.10.: Geographic Maps of Spatiotemporal Model RMSE across data regimes. Many sites maintain similarly low RMSE with a few larger outliers. The worst site is the same across the data regimes. The inclusion of PCs changes the worst-perfroming site.

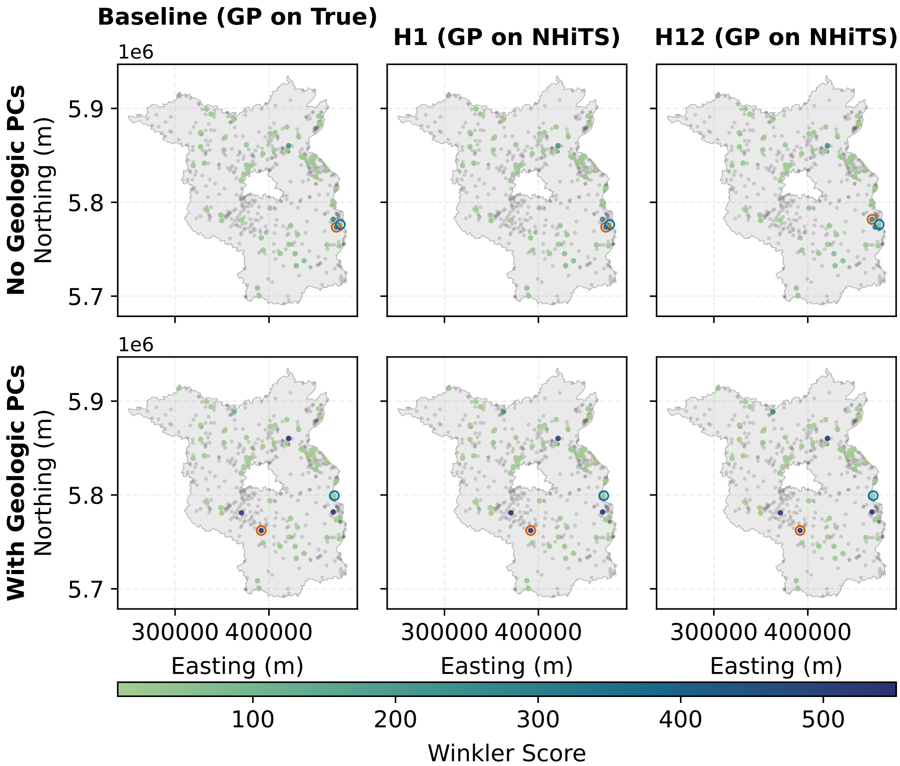

Figure 4.11.: Geographic Maps of Spatiotemporal Model Winkler Scores. The map is dominated by a few poor-performing outliers. The GP without PCs has very little geographic distance between the best and worst performing sites, indicating that geographic space does not inform model interval predictions well. When given PCs, the best and worst sites spread though there are a few outliers.

4.5.2. Performances in Principal Component Space

The performances in Principal Component space are analyzed in order to address the complexity of the geology in geographically nearby locations. That is, there are locations which are nearby and exhibit very different geological properties. We select the principal components 1, 2, and 8 due to the proportion of variance they explain (1 and 2) and the feature importance from the GP lengthscale analysis in prior sections (8). As covered in Section 3.2, there are meaningful interpretations that can be drawn for each of the principal component axes based on their loading vectors. We summarize them here:

- PC1: Higher values indicate lower Topographic Wetness Index (TWI).
- PC2: Higher values indicate higher stream distance.
- PC8: Higher values indicate higher elevation and lower groundwater recharge rate.

Additionally, the metrics examined in principal component space are both the RMSE and the Winkler Score to gauge the deviation from target and distributional forecasting performance, respectively. We begin with the RMSE in principal component 1 and 2 space.

RMSE in Principal Component Space The following figures examine the RMSE in principal component space to investigate the geologic factors behind large deviances from the true values.

In Figure 4.12 one can see in all plots that wells with large values of PC1 have lower performance relative to the others. Additionally, the worst performing well in the lower row of plots, with geologic principal components, has the largest values of PC2. In all cases, the best performing wells are near other wells in PC space, indicating geologic similarity with other testing wells. There are some anomalies, for example in the case of the best performing well in the bottom rightmost plot, the two wells closest to it perform poorly in comparison.

In Figure 4.13, one can see there is less deviation on the PC8 axis between the best and worst sites in the models without geologic PCs than in the models with geologic PCs. Additionally, there is agreement in each row of the worst performing sites, all having large values of PC1. In both this figure and the last, the performances when PC1 is smaller tend to be better. Finally, as in the last figure, the best performing wells tend to be near other wells whereas the worst performing wells tend to be near to the extrema of the axes.

Winkler Score in Principal Component Space The following figures examine the Winkler score in Principal Component space to facilitate investigation of the geologic drivers behind high and low distributional forecast performance.

In Figure 4.14, there a few extreme outliers in the case of the GP fit with geologic principal components. These outliers, which indicate low distributional forecast performance, tend to have large values of PC1 and occupy positions in the space typically farther from the other points. Specifically, the worst performing site in the case of the GP with geologic PCs, which is the same across the columns, has maximal values of PC1 and PC2 of the other wells in the test set. In the case of the models without geologic principal components, the worst performing sites all have larger values of PC1 than the best performing sites.

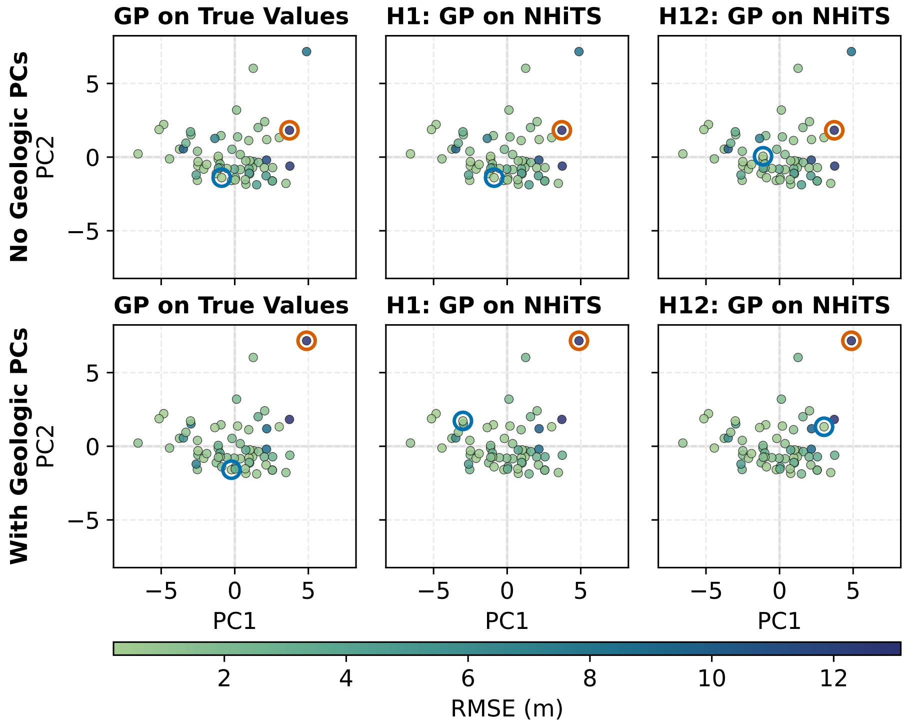

Figure 4.12.: RMSE in PC1-PC2 Space. In all scenarios, the worst perfroming sites have large values of PC1 and PC2. The worst site is the same between the data regimes of the GP with only coordinates. The similarity holds for the GP with PCs as well.

In Figure 4.15 we see similar results with the main differentiatior in all plots being the values of PC1. In all plots, including the models without geologic PCs there are differences in the PC8 values between the worst and best performing sites with the worst performing sites having higher values of PC8. In the case of the models containing geologic PCs, the worst performing sites all have the largest values of PC1 in the test set.

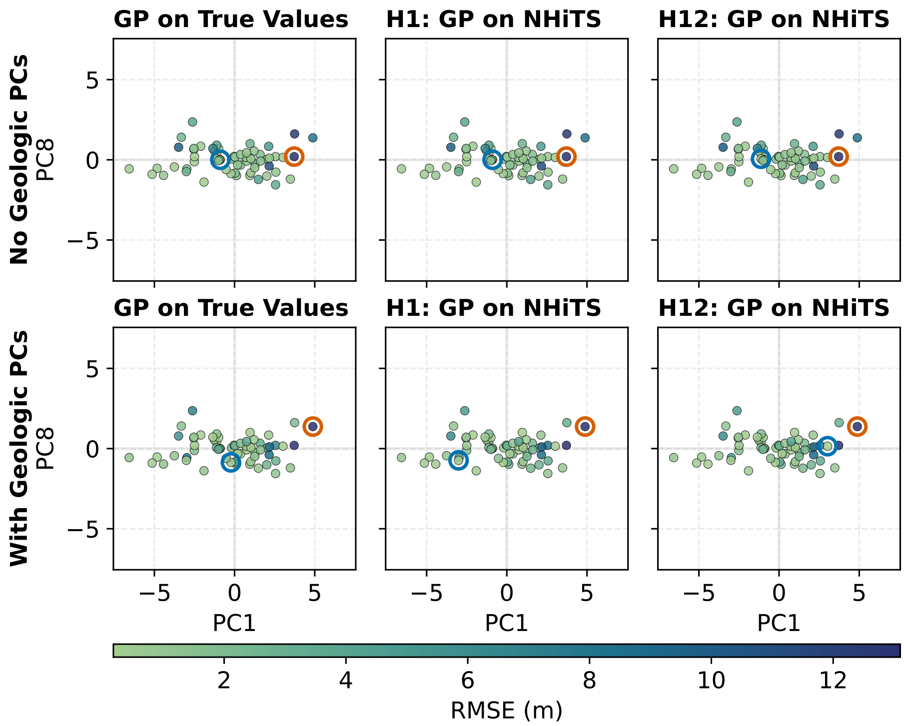

Figure 4.13.: RMSE in PC1-PC8 Space. The worst performing sites tend to have larger values of PC1 across all plots. Without PCs, there is little variation in the PC8 axis between the best and worst performing sites. In the models with PCs, the worst performing sites have larger values of PC8.

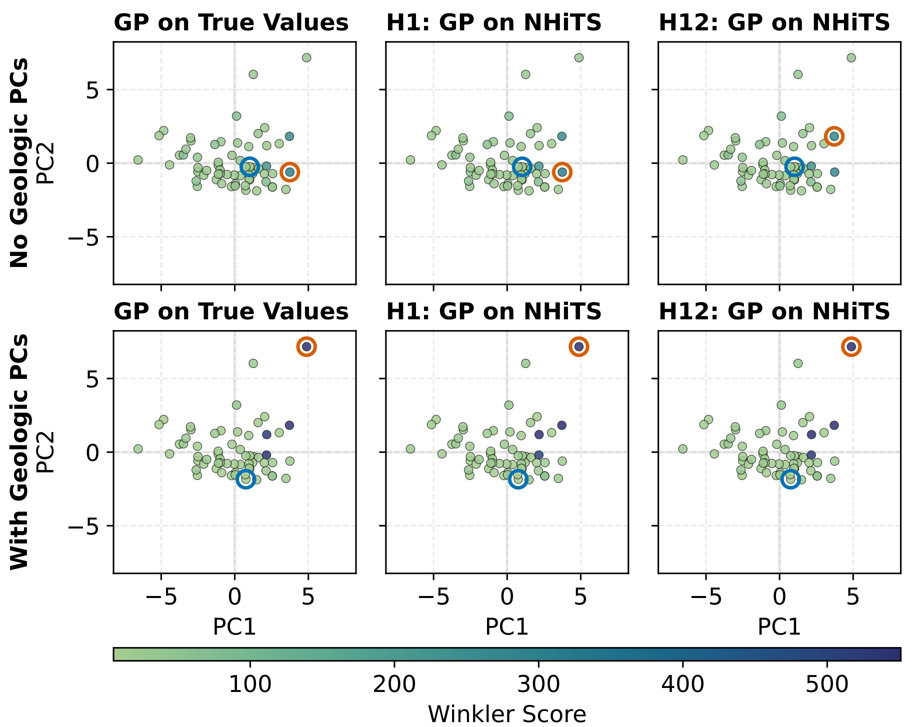

Figure 4.14.: Winkler in PC1-PC2 Space. Across all plots, the worst performing sites have larger values of PC1 than the best performing sites. Without PCs, the worst performing and best perfroming sites have similar PC2 values. With PCs, there are larger-valued outliers and the worst perfroming sites have larger PC2 values.

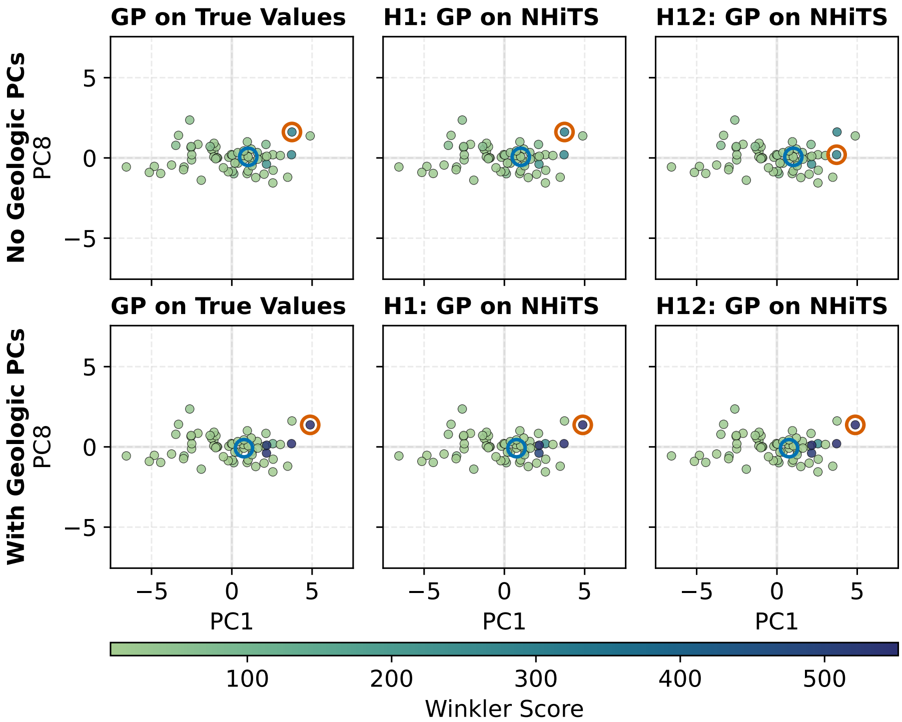

Figure 4.15.: Winkler in PC1-PC8 Space. The worst performing sites again tend to have larger values of PC1 than the best perfroming. There is a less pronounced change in difference of PC8 values for the best and worst performing sites in the models with and without PCs.

4.5.3. Micro Analysis

This section is concerned with better understanding the value of some metrics seen earlier, specifically, the RMSE and the NSE. To do so, the best, median, and worst forecasts based on the evaluation metrics in a scenario described below are visualized along with the prediction intervals and the true values. The scenarios we examine are the best, median, and worst forecasts of the GP with geologic principal components and on horizon 1 N-HiTS forecasts. The corresponding graphics are the center-bottom panes of the previous plots in geographic coordinate space and principal component space. The reasons for this choice are the large difference in the PC1 and PC2 values, large geographic distance, the forecasts are from the selected *best* decoupled spatiotemporal forecaster, and the fact that it is a spatiotemporally forecasted value, not one from GPs fit on true values.

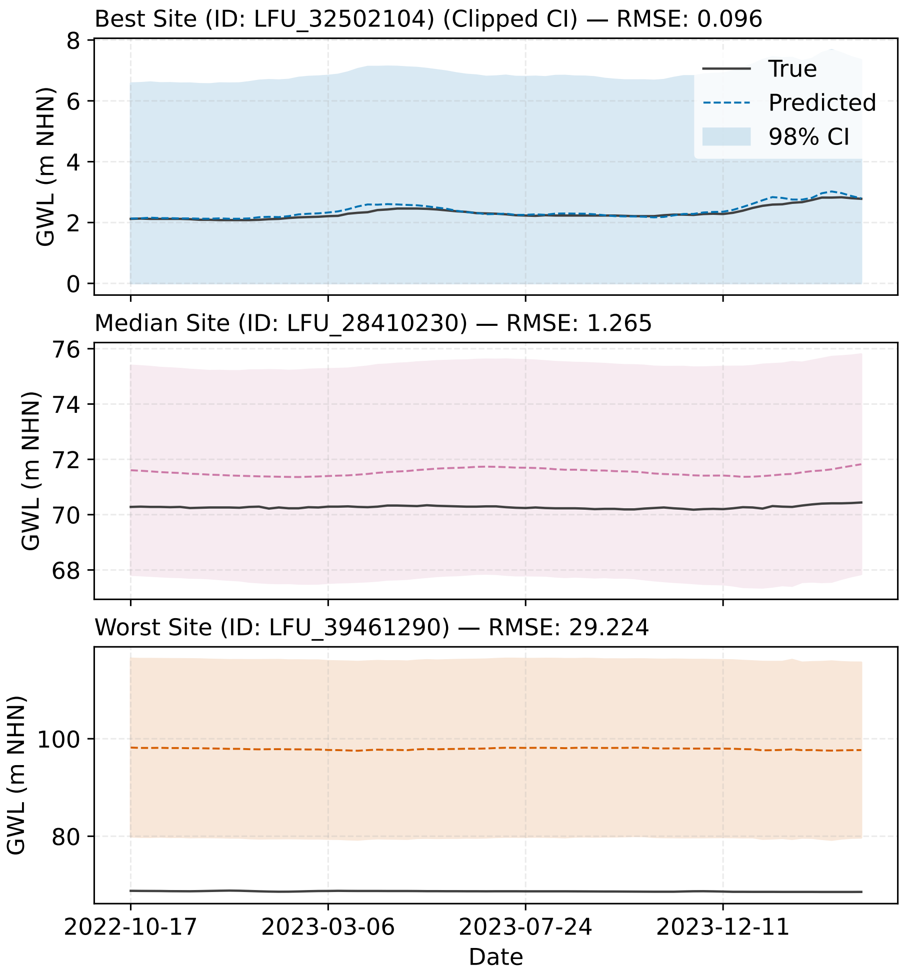

Figure 4.16.: Best, Median, and Worst Forecasts w.r.t. RMSE from H1 GP with Geologic Principal Components on N-HiTS Forecasts. The worst forecast overestimates the true series by about 30 m NHN and has a much wider interval.

In Figure 4.17 we visualize the best, median, and worst forecasts with respect to NSE in the same modeling setup to better characterize the metric values. Given that we examine the median, appoximately 50% of sites will have performance between that seen in the best and median plots and approximately 50% of sites will have performance between that seen in the median and the worst plots. This allows for characterization of the forecast performance visually relative to the metrics provided in the previous sections.

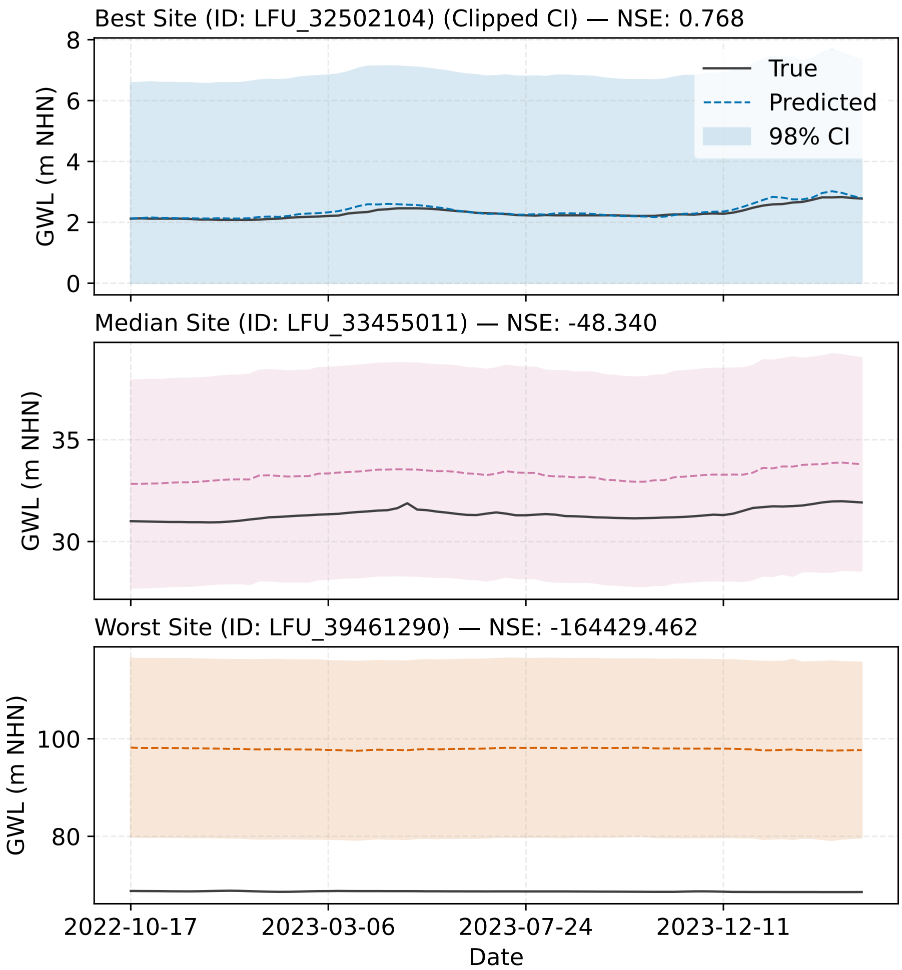

Figure 4.17.: Best, Median, and Worst Forecasts w.r.t. NSE from H1 GP with Geologic Principal Components on N-HiTS Forecasts. Even though the NSE is -48.34, the median plot misses the true value by less than 3 m NHN throughout.

A next step would be to determine when the forecasts are more similar to the best and more similar to the worst and if there is a highly correlated value or predictor of this. Preliminary analysis is covered in Appendix D.4.2 but more work could be done as is outlined in the future work section.

4.5.4. Relative Comparisons

In this section, we compare the performance between the Gaussian processes fit on historical known-site means, true values, and forecasts of the two horizons. We again look at the best performing model on the validation set, the Gaussian process with geologic principal components fit on N-HiTS forecasts. Each plot depicts the metric value of the two methods at each site so that coordinates of the form (baseline, target) are plotted. The metrics examined are RMSE, NSE, and Winkler score so that we can better understand the deviation from the true value, the performance in a standard hydrology metric, and the model uncertainty calibration. All metrics are computed on the Test-Test set. The plots omit outliers defined using the boxplot definition, non-outliers are in the set: $[Q_1 - 1.5 \cdot \text{IQR}, Q_3 + 1.5 \cdot \text{IQR}]$ The plots depict the values for each metric of the baseline Gaussian process fit on true values or site means on the x-axis and the target, the Gaussian process fit on N-HiTS forecasts on the y-axis. Finally, comparisons with the baseline Gaussian processes not using principal components are included in Appendix D.4.3.

Horizon 1 We begin the analysis on the horizon 1 forecasts compared to the baselines.

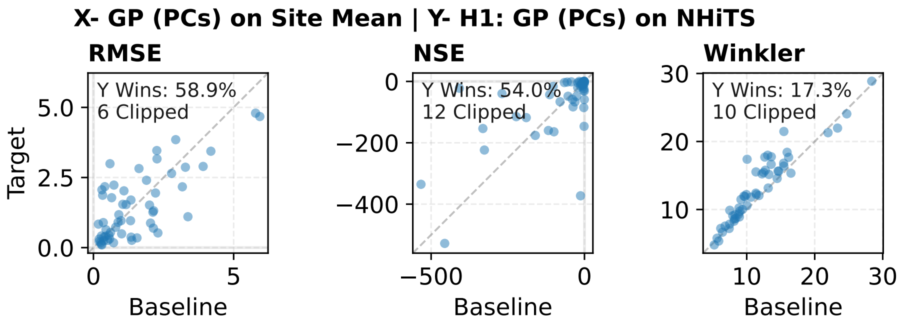

Figure 4.18.: Relative Performance of GP on Means and GP on N-HiTS H1. The GP on N-HiTS H1 model wins in the majority of cases in RMSE and NSE and loses in the majority of the Winkler cases.

In Figure 4.18, the GP on N-HiTS forecasts outperforms the GP on historical means in RMSE and NSE in the majority of wells and underperforms in the Winkler score. In the RMSE plot, it appears that the high-value points have slightly smaller errors when the GP is fit on forecasts. The NSE has highly negative values though the distributions have long tails and many values are near 0. Finally, the Winkler score has similar performance for many of the wells though the baseline outperforms in the majority of the wells.

In Figure 4.19 the GP on true values is compared to the GP on N-HiTS forecasts. In this case, the GP on forecasts does not outperform in the majority of wells for any metric though it ties the GP on true values in RMSE. There appears to be an even spread of the values in the RMSE plot and underperformance in NSE. The Winkler score plot depicts slightly higher values for the GP on forecasts than the GP on true values.

Horizon 12 We now examine the horizon 12 forecasts relative to the baselines.

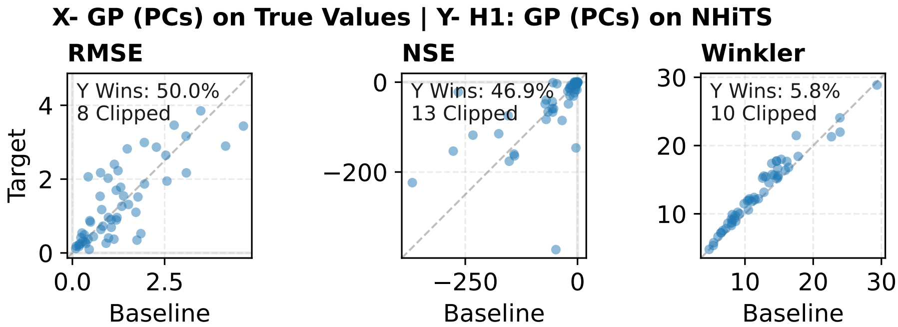

Figure 4.19.: Relative Performance of GP on True Values and GP on N-HiTS H1. The GP on N-HiTS H1 model loses in the majority of cases in NSE and Winkler but maintains a 50-50 win rate in RMSE.

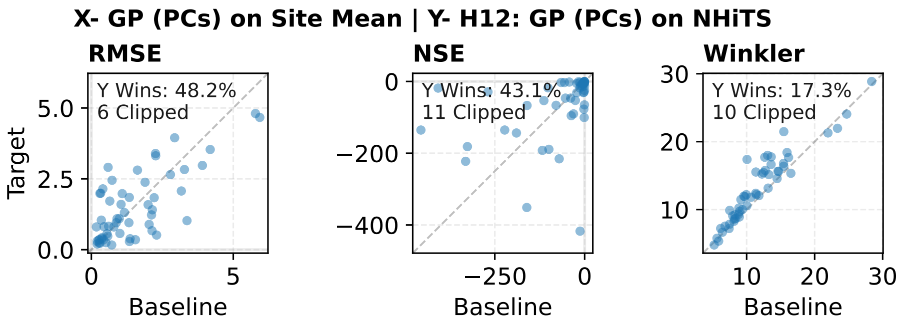

Figure 4.20.: Relative Performance of GP on Means and GP on N-HiTS H12. The GP on N-HiTS H12 model loses in the majority of cases in RMSE, NSE, and Winkler. The RMSE is 1.8% from equality.

In Figure 4.20 we observe similar behavior to Figure 4.18 though more underpreformance of the GP fit on N-HiTS forecasts. The RMSE and NSE plots have approximately 10% and 11% fewer wells, respectively where the GP on forecasts outperforms the GP on site means when compared to the GP on horizon 1 forecasts. Finally, the Winkler Score has the same proportion of wells where the GP on forecasts outperform the GP on site means and maintains a similar shape of distribution as in the horizon 1 scenario.

In Figure 4.21 also exhibits a degredation in performance of the GP fit on N-HiTS forecasts. The RMSE plot has approximately a 5.6% decrease in the proprotion of wells where the GP on forecasts outperforms the GP fit on true values. Additionally, the NSE has a decrease of outperformance proportion of approximately 4.6%. Finally, the Winkler score sees an increase in the performance from the horizon 1 GP on N-HiTS forecasts of 1.9%.

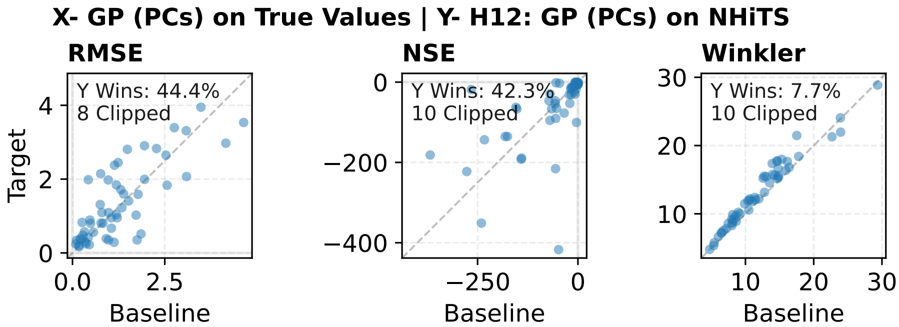

Figure 4.21.: Relative Performance of GP on True Values and GP on N-HiTS H12. The GP on N-HiTS H12 model loses in the majority of cases in all metrics. The RMSE is 5.6% worse in terms of win rate and the Winkler score, while losing does not appear to be substantially lower in many individual cases.

4.6. Summary

This chapter presented the results from the methods outlined in Chapter 3. It documented model performances and optimal hyperparameters from methods of each of the tasks of temporal forecasting, spatial interpolation, and spatiotemporal forecasting. The first task, temporal forecasting, evaluated several neural forecasting models in addition to a simple baseline method and utilized a suite of evaluation metrics to gauge the performances of the methods relative to one another and to the data. Of the models examined, the N-HiTS and Chronos2 Finetuned models performed best across the problem space. The spatial interpolation task examined the performance of various model classes including tabular foundation models, deep neural networks, Gaussian processes, and a simple nearest neighbor baseline in predicting the average groundwater level at unseen locations. The highest performing models were Gaussian processes. The spatiotemporal forecasting task was split by model design into a coupled and decoupled section where the former jointly trained models to do both forecasting and spatial interpolation and the latter first made predictions with a temporal forecaster and did spatial interpolation on these forecasts. Of the models in the decoupled task considered, the Gaussian processes with geologic principal components fit on N-HiTS forecasts maintained high relative performance across the metrics. In the case of the coupled models, the cross attention models outperformed the Kronecker Gaussian process in point prediction but underperformed when the distributional metrics are considered. The Gaussian process with geologic principal components on N-HiTS forecasts outperformed the other spatiotemporal models considered which motivated further analysis of its performances. Simultaneously to the evaluation of the model performances, efforts towards model explainability with respect to feature importance were undertaken in the case of the Gaussian process models with the analysis of lengthscales used as a measure of importance.

The final section in the results chapter consisted of deeper analysis of the high and low performances in the Gaussian process on N-HiTS. To this end, geographic maps with point forecast and distributional forecast metrics are considered to investigate patterns within the geographic distribution of high and low performing sites. Additionally, to address the geographic complexity not provided by geographic maps, analysis of model performances in principal component space allows one to identify patterns of performances relative to the covariates informing the spatial model. To better understand the forecasting performance, the best, median, and worst performing forecasts relative to the RMSE and NSE were plotted. Finally, relative comparisons are made to explore the degradation of the spatiotemporal forecasts when the spatial model is given true values instead of forecasts at known sites. The relative comparisons also examine the differences between Gaussian processes fit on historical site means at known sites and those fit on the forecasts of models given the context of known sites.

5. Discussion

This section explores methodology and results associated with each of the tasks considered in the thesis.

5.1. Temporal

The temporal section covered the task of forecasting future groundwater levels at sites with known historical groundwater levels. Methods applied to this task were Chronos2 Zero-Shot, Chronos2 Finetuned, N-HiTS, and the Naive Baseline. The predictions at horizons of 1 and 12 weeks were compared to the true values at each site using RMSE, NSE, KGE, MPIW, PICP, and Winkler Scores.

The results recorded the optimal hyperparameters for each of the models and the performances of the metrics on the validation and test sets at each horizon. The N-HiTS and Chronos2 Finetuned models outperformed the other models in all metrics with a larger increase in performance over the baseline in the horizon 12 forecasts. The N-HiTS model had the best performance on the validation and test sets of all other models and was chosen as the forecaster for the decoupled spatiotemporal models.

The errors in RMSE, for ease of comparability between sections, were far smaller in the temporal section than in the spatial section or the spatiotemporal sections. The RMSE and NSE of the temporal section were similar to the global models explored in a motivating work, Kunz et al. [27]. Additionally, the PICP was reasonably high with small MPIW in the Chronos2 Finetuned and N-HiTS models indicating good calibration of the distributional forecasts. The main findings of this section were:

- The Chronos2 Zero-Shot model is less effective than the naive baseline in the horizon 1 but more effective in horizon 12 and provides distributional forecasts in both.
- The N-HiTS model performs best in horizon 1 and 12 of the models considered.
- The Chronos2 Finetuned model is on par with the performance of N-HiTS at horizon 12 and still outperforms the baseline at horizon 1.

5.2. Spatial

The spatial section covered the task of interpolating the mean groundwater level at sites with unknown groundwater levels, evaluated by means of spatial validation and test sets. Methods applied to this task were, K-NN, TabPFN, ResNet, and Gaussian processes. The predictions were compared to the mean of the groundwater level over time at the holdout sites and metrics reported include the spatial RMSE, spatial NSE, spatial MAE, MPIW, PICP, and Winkler Scores. Each site had one mean groundwater level and one prediction, making the problem a point prediction task rather than forecasting task.

The results recorded the optimal hyperparameters for each of the models and the performance metrics on the validation and test sets. The best performing model on the validation set was the ResNet while the Gaussian process with only coordinates performed better on the test set with respect to both RMSE and NSE. The ResNet had lower PICP on both the validation and test sets while the Gaussian processes maintained higher PICP in all cases, presumably due to the lower MPIW in the case of the ResNet and the higher MPIW in the case of the Gaussian processes. The interpretability of the Gaussian process length scales also allowed for some interpretation of feature importances to be done. Notably, the geographic coordinates were most important with the geologic principal components 1 and 8 showing the highest importance of all covariates, a result seen in the decoupled section as well.

The RMSE in the spatial section was higher than the RMSE in any other section. The ResNet had the lowest error of all models on the validation set which indicated the potential for high performance in the spatiotemporal cross attention model as it would be able to get the offset correct. This ordering swapped with the simpler Gaussian processes performing better on the test set. The Gaussian process models attained higher performance on the test set than the validation set, and the variant without a covariate informed kernel performed best on the test set in terms of NSE and RMSE. The Gaussian processes fit here were later used in the relative comparisons against the Gaussian processes fit on forecasts as a baseline. The main findings of this section are:

- All spatial interpolation methods have more error in RMSE and larger prediction intervals on average than the temporal methods.
- The geographic coordinates and geologic principal components 1 and 8 are most important in the Gaussian process models.
- There is a high variance in the spatial dimension of the data, illustrated by the reordering of the best performing models between the validation and the test sets.

5.3. Decoupled Spatiotemporal

The decoupled spatiotemporal section covered the task of forecasting at sites with no historical groundwater levels using a two step pipeline of forecasting and then interpolating these forecasts. This was evaluated by means of a spatiotemporal holdout set with values in the future at unseen locations relative to the data models were trained on. The forecasters used for this were N-HiTS, Chronos2 Zero-Shot, Chronos2 Finetuned, and the naive baseline. The spatial interpolators used were Gaussian processes with varying sets of features given to the kernels. The interpolations of forecasts were treated as forecasts at the heldout sites where the true values were compared in a manner identical to that of the temporal forecasting section at horizons of 1 and 12 weeks.

The coordinate GP on forecasts (without covariate-informed kernels) has the lowest median RMSE across sites of all models tested and little difference between the forecast horizons on the test set. The geologic principal component informed GP has slightly higher median RMSE but better distributional forecasts with the same coverage proportion (100%) and a smaller MPIW. Additionally the geologic principal component GP had better validation set performance. The GP lengthscales were similar to the spatial section though aggregated over the numerous GPs fit in this section. The lower fence of the GeoTessera principal component lengthscales distribution boxplot had a higher value than the upper fence of three geologic principal components and the coordinates, indicating lower importance relative to these geologic principal components and coordinates.

Relative to the coupled spatiotemporal modeling section, we see high RMSE performance in the coordinate GP on forecasts on the test set. Additionally, while wide on average, the prediction intervals cover the true point more often, shown by the higher PICP. This could be attributed to the simplicity of the models, indicating the potential of overfitting and the preference of a lower bias model in the task examined. We summarize the main findings here:

- The coordinate GP on forecasts model performed better than the other GP variants and the coupled spatiotemporal models on the test set.
- The GP with geologic prinicpal component information on N-HiTS had high performance in terms of RMSE and NSE while maintaining high KGE on the validation set.
- The feature importance rankings remained mostly the same as in the spatial section with the geographic coordinates most important followed by PC 1 and 8.
- The forecast horizon did not appear to have a large effect on the overall predictive accuracy, though this could be attributed to the scale of the spatial error relative to the temporal error and the difference in data density.

5.4. Coupled Spatiotemporal

The coupled spatiotemporal section covered the task of forecasting at sites with no historical groundwater levels using a joint pipeline that simultaneously forecasts in time and interpolates in space. The two main classes of models evaluated are: The Kronecker Gaussian process which directly models the groundwater level given spatial and temporal information and the cross attention model that uses pretrained neural forecasting and spatial models to fuse the representations of the inputs using attention before passing into a quantile forecast head or variational Gaussian process. This task was evaluated using the same spatiotemporal holdout set with values in the future at unseen locations relative to the training data of the models. The forecaster used in the attention models was N-HiTS and the spatial model used was the ResNet, both due to the compatability of the underlying PyTorch software framework and the high validation set performances. The attention models both made forecasts at horizons of 1 and 12 weeks similar to the previous models. The Kronecker Gaussian process treated the spatiotemporal forecasting differently where no horizon was used and any point in space and time could be inferenced for a prediction. The metrics used are the same as the other tasks.

The primary comparison of the Kronecker Gaussian processes was to understand the benefit of using the geologic principal components. When supplied with the geologic covariates, the performance was on average higher with better interval predictions characterized by a smaller MPIW, the same PICP of 100%, and a lower RMSE, NSE, and KGE. The hyperparameters indicated a much larger importance of the spatial features relative to the temporal index, which remains consistent with the errors between the two dimensions of the task we have observed to this point. The attention models attained lower RMSE but far worse coverage than the Kronecker Gaussian processes, shown by the PICP with a median near 0% for both attention models. The attention models had a degredation in all metrics except KGE as the forecast horizon increased. Still, the KGE was negative indicating poor forecasting performance. Additionally, both neural models utilized dropout with an optimal value near the search space upper limit of 0.4. This could indicate that there was overfitting present but confirmation would require further investigation.

The coupled section generally had results worse than the decoupled section on the test set. This was unexpected but could make sense if the bias variance tradeoff and data scarcity of the spatial problem are considered. Particularly, the models in the coupled section remove some inductive bias as far as the independence of the behaviors of the temporal and spatial aspects of the system, exacerbated by the difference in sampling sufficiency between the two dimensions of the problem. That is, the decoupled models have more bias by design and are therefore less prone to overfitting, especially when spatial data are scarce. Further development of neural approaches to this problem with a specific focus on remedying overfitting could be explored but we leave this as future work. The main findings of this section are:

- The temporal index had less importance than the spatial coordinates in the Kronecker GP, supporting the results of previous sections.

- The geologic principal components increase the interval prediction performance of the Kronecker GP model but fail to provide substantial benefit to the point predictions on the test set, as measured by the metrics considered.

- The attention models both underperformed relative to the decoupled models on the test set and had large optimized values for the dropout hyperparameter, suggesting the presence of overfitting.

5.5. Best Spatiotemporal Models & Relative Comparisons

The model chosen for further analysis was the geologic principal components and geographic coordinates GP on N-HiTS forecasts due to the high validation set performance across both point and interval forecast metrics, the relative simplicity of the model, the ability for interpretation of the covariates through the use of Gaussian process lengthscales, and tighter connection to the principal component biplots. The locations of extreme forecast performance (min/max) in geographic and principal component space were identified to investigate patterns of failures and successes. Thereafter, an analysis of the performance relative to an upper baseline, Gaussian processes fit on true values, and a lower baseline, Gaussian processes fit on observed, historical site means is done to evaluate the differences in performance when interpolating forecasts and fitting on known data.

5.5.1. Geographic Maps

The analysis of the geographic distribution of site performances consists of a comparison between GPs with and without geologic principal component kernels and context labels of the true values at the known sites, the horizon 1, and the horizon 12 forecasts at each time. The main findings are as follows:

- Poor performing sites do not necessarily occur in data scarce regions - the geographic distance between the context points and poor performers is sometimes lower than that of the high performers.
- Poor performing sites may be collocated with higher performing sites in a nearby geographic neighborhood - there are no clear clusters.
- The poor performing sites maintain consistency between the GPs on true values, GPs on horizon 1 forecasts, and GPs on horizon 12 forecasts.
- The Winkler score contains outliers in the case of GPs with geologic principal components despite having better median performance than the GPs only on coordinates in the decoupled section, indicating large failures occasionally occur.

5.5.2. Principal Component Space

The analyses of site performances in principal component space use a similar setup to the geographic analysis but seek to better understand the behavior of the spatiotemporal models in "geologic space" such that the aspects of the environment not shown in the geographic coordinates can be explored. The main findings are as follows:

- The worst performing sites had low values of topographic wetness index.
- The worst performing sites were the same between the context groundwater levels (true, H1, H12) but changed if the GP itself was provided geologic PCs.
- The best performing sites varied between the data the GPs were fit on, with different sites for the GP on true values, GP on H1 forecasts and GP on H12 forecasts.
- When provided with geologic PCs, the best performing wells tended to have lower values of the stream density variables.
- Generally, the worst performing sites had extreme values in principal component space, indicating predictions made on out-of-distribution spatial data are challenging for the models.

5.5.3. Micro Analysis

The micro analysis visualizes the forecasts, true values, and forecast intervals of a selected spatiotemporal model to better understand the best, median, and worst performing forecasts relative to the evaluation metrics. The main findings are:

- The best performing sites in RMSE and NSE tend to follow the true series well with a small forecast interval.
- The median sites in RMSE and NSE tend to follow the true series well with a larger offset than the best performing site, and still maintain a tight interval, containing the true series.
- The worst performing sites forecast values far away from the true series with a wide interval not containing the true series.
- The variance and temporal correlation of the worst performing forecasts appear similar to that of the true series, indicating the low performance could be due to the spatial interpolation not the forecasting.

5.5.4. Relative Comparisons

The relative comparisons seek to examine the effect of fitting GPs on forecasts of known sites, historical means of known sites, and true values of known sites. They compare the GP on forecasts to the theoretical lower bound of the GP fit on historical known site means and the theoretical upper bound of the GP fit on true values at the same timesteps as the forecasts. The main findings are:

- The horizon 1 GP outperforms the GP on means but underperforms the GP on true values for RMSE and NSE and underperforms both in the majority of cases for Winkler.
- The horizon 12 GP underperforms the GP on means and GP on true values in the majority of cases for all metrics.
- More investigation with in-production spatial models should be conducted to examine the performance improvements of the decoupled spatiotemporal approach.

6. Conclusion

In this chapter, we close the thesis with a discussion of the methods, results, shortcomings, and future directions. We begin by summarizing the previous chapters, the contributions, and the discoveries encountered within this work. Thereafter, we address the core questions outlined in Section 1.2, seeking to provide answers using the information uncovered during the thesis. Finally, we cover limitations and future work.

This work was concerned with the spatiotemporal modeling of groundwater levels in Brandenburg, Germany. To this end, the task was formalized and separated into separate problems of temporal forecasting, spatial interpolation, and spatiotemporal forecasting. Related work was explored in the hydrological, machine learning, data science, and hydrological machine learning communities. In addition to the literature, collaborators from the German Federal Ministry of Natural Resources were consulted for advice, the data, and direction.

The data were explored and characterized, principal components analysis was conducted to better visualize and model the higher-dimensional covariate space, and rigorous data splits in both space and time were carried out to ensure the generalizability of the model was effectively measured with the metrics. The methodology for each of the tasks were clarified with the specification of hyperparameters as needed. Performance evaluation metrics and a description of the training setting were expanded upon before moving into documentation of the results.

The results of each task were recorded and visualized for each model and each metric examined. The optimal hyperparameters from each search space were recorded for each model, accounting for differences in model paradigm and methodologies. An examination of the best performing model including consideration of covariates which correlated with loss metrics was conducted through geographic maps and plots of principal component space. Select forecasts were visualized along with the true series and the forecast intervals. Finally, relative comparisons between baseline models and the best performing spatiotemporal model were made.

The results of these analyses are manyfold, importantly, a better understanding of the performance of spatiotemporal forecasting on the considered dataset with neural forecasting models and Gaussian processes was gathered. Additionally, the forecasting performance of several neural forecasters including the zero-shot capable Chronos2 were compared on this dataset. The results of Kunz et al. [27] were replicated, further confirming their findings. The effectiveness of TabPFN on this dataset relative to other models such as K-NN, Gaussian processes and a novel ResNet design was documented. The relationship between static covariates and spatial model performance was examined, providing insight into geologic environments within Brandenburg that may be better suited for deployment of such models. Finally, the examination of Gaussian process performance degradation when fit on forecasts instead of true values is examined.

6.1. Goals of the Research

In the Introduction, five core questions were mentioned. They are addressed here with the findings of the thesis as support for the answers.

1. Can foundation models (zero-shot or finetuned) be used for groundwater level forecasting?

In the temporal section of the thesis, we found that the results were competitive with the Chronos2 finetuned model, only slightly below the best model, N-HiTS, across the metrics. In the 12 week horizon case, the Chronos2 zero-shot outperforms the naive baseline and offers distributional forecasts. Due to these points, time series foundation models are reasonably effective at forecasting the groundwater level and, in broader scope, could be worth considering further.

2. Can decoupled spatiotemporal models with forecast-interpolate steps effectively forecast the groundwater level at unmonitored sites?

In the decoupled spatiotemporal section of the thesis, we investigate this question. The answer is not completely clear due to the poor performance relative to the purely temporal section. In order to better evaluate the forecasting performance relative to the limitations of the spatial models, relative comparisons are made where the spatial models are fit on true values as an *oracle forecast* and on historical means as a simple, long-term forecaster. The degredation of the spatial models when fit on forecasts as opposed to true values is moderate and the models fit on forecasts outperform the models fit on historical means in some cases. This suggests the decoupled approach could function as a forecaster at sites with unknown historical groundwater levels with the caveat that performance is limited by the spatial model.

3. Do coupled spatiotemporal models with joint training of the spatial and temporal aspects outperform decoupled models in groundwater level forecasting?

As shown in the decoupled and coupled spatiotemporal sections, the performance of the Gaussian processes fit on forecasts performs better than the jointly trained, coupled models on the test set in many cases. We explored only a few jointly trained models which limits the reach of this claim.

4. Do geologic covariates improve the predictive performance?

In the spatial and decoupled spatiotemporal sections we investigate this point. In the spatial section, on the validation sets, the geologic principal components Gaussian process outperforms the Gaussian processes without them but on the test sets, there is worse performance in RMSE and NSE with the geologic principal components. Additionally, the MPIW and Winkler scores are lower and the PICP remains similar when there are geologic principal components than without on the validation set, indicating the model is more confident with the principal components and still maintains high coverage. There is also less spread in the values on the test set of the decoupled spatiotemporal when geologic principal components are provided, indicating more consistent performance across timesteps and horizons. Due to these mixed results, it is inconclusive as to whether the geologic principal components help in performance. While they do appear to reduce the variability of the model and reduce the predicted standard deviations, they do not necessarily help with the point prediction/forecast ability in all cases, further investigation of this point may be required.

5. Which geologic covariates correlate with higher/lower performance?

Using the figures depicting model performance in principal component space, one can see that points towards the extremes of the axes have lower performance. While this is likely due to the difficulties associated with models predicting out of distribution samples, the asymmetry in principal component space indicates that there could be an influence of the covariates on which the principal components analysis was conducted. Specifically, larger values of PC1 correlate with lower performance in RMSE and the loading vector for topographic wetness index variables goes along the negative PC1 axis. This indicates that sites with lower values of topographic wetness index may have worse performance relative to those with hgiher values of topographic wetness index. A similar observation can be made with the PC2 axis where the loading vectors associated with stream distance follow the PC2 axis in the positive direction. There is a correlation between larger values of PC2 and lower performance in RMSE which indicates that sites with larger stream distance may have lower performance. Consultation with professionals at the German Federal Ministry of Natural Resources and these findings suggest these geological factors could play a role in the model performance.

6.2. Limitations

This section outlines notable limitations of model performance and methodology and provides possible explanations where relevant.

6.2.1. Performance

A main shortcoming of the spatiotemporal forecasting models examined when compared to the temporal forecasting models was the degredation in performance. When the historical levels of the groundwater are known, the temporal model must only learn fluctuations which can be recovered when the models account for seasonality, trend, and covariates. Moreover, the performance of the naive baseline shows that there are only minor performance gains to be had when a more advanced model is used. When the model must instead forecast at sites with no known historical groundwater levels, the model must also learn the offset which, as seen in the spatial task corresponds with larger errors than temporal forecasting. Consequently, this results in a much larger deviation from the true values of the series at the holdout wells due to the poor acquisition of the offset. Succinctly, the error in the spatiotemporal models is a combination of the spatial interpolation and the temporal forecasting errors where the spatial interpolation tends to deviate farther from the targets on average.

As far as causes, the sampling density of the data relative to the variance in the area sampled is likely a driving factor for this contrast in performance between the spatial and temporal methods. Another reason for the poor performance of the spatial task when compared to the temporal task could be the lack of necessary covariates. The data used in these analyses were broad in scope consisting of geologic covariates selected by geologic subject matter experts and remote sensing data through GeoTessera, but perhaps there are unutilized covariates such as anthropogenic factors that would have made the models more performant while not leaking target data into the model.

6.2.2. Methodological

A methodological limitation could have been spatial size of the study area. As expanded upon in the previous section data density appeared to be a persistent issue in the spatial performance and consequently the spatiotemporal performance of the models examined. To this end, if we had decided to focus the study area on a smaller geographic region, perhaps with more geological homogeneity and denser spatial data sampling, the performance may have been better. As seen in Appendix A, there are regions in Germany with a higher spatial density of wells such as the city of Berlin and the regions along the Rhine.

As far as the models implemented, a limitation was the performance of the coupled spatiotemporal models. While these models were expected to perform better, they ended up performing worse in some cases than the simpler decoupled spatiotemporal models. Additionally, the Chronos2 foundation model was expected to perform better in the forecasting task than the N-HiTS model but the results show otherwise. A possible cause of both of these underperformances could be related to the bias-variance tradeoff and the simplicity of the opposing models.

We could have examined more spatial models in the decoupled spatiotemporal modeling such as K-NN but the higher performance, probabalistic outputs, and strong support of the Gaussian processes in the hydrological community made this direction seem less relevant. More folds in the spatial holdout sets could have benefited the analysis. As a larger project, and beyond the scope of this Master's thesis, more information on the operational setting in which these models are deployed and the exact processes where integration of the forecast models with spatial models could be deployed could have been beneficial in directing the project towards an outcome with maximal operational effect. Finally, the consideration of the spatial boundaries of the study region was not explored. In order to have an operationally effective model, a better understanding of when the spatial interpolation ability substantially degrades could be of interest.

6.3. Future Work

We cover several directions for future work, beyond this thesis. We believe these points will help to address the shortcomings outlined in the limitations section and could provide useful directions for others interested in continuing the work.

6.3.1. Scientific Directions

Through this thesis, we have found that the decoupled models are promising in terms of predictive performance and specifically the Gaussian process fit on forecasts did not degrade substantially in performance relative to the Gaussian process on true values in all cases. The Gaussian process on forecasts also outperformed the Gaussian process fit on historical site means in some scenarios. To this end, a future direction for scientific work could be to devote more efforts specifically to evaluating the exact current state of the art Gaussian process models used by the BGR and simply switching out the true values with the forecasts produced by a performant neural forecaster. This would also require careful attention to the problem of evaluating model generalization such that the evaluation sites remain isolated in space and time from the model during training to ensure the performance is effectively compared.

Additionally, more work on predicting when the models developed can be trusted, fail, and perform well could be done. This could consist of creating meta-learners that could give a reliable prediciton of when the model predictions are operationally viable.

Finally, to address the spatial aspect, the creation of another pipeline step that makes the spatial interpolation task easier could be beneficial. Specfically, if a clustering algorithm was used in the spatial covariate space, separate Gaussian processes could be fit per cluster, horizon, and timestep instead of just horizon and timestep, ensuring that the interpolation happens between physically similar locations and potentially avoids spatially out of distribution predictions.

6.3.2. Acquisition of Additional Data

A main finding in the thesis was that the difference in the scale of the errors between the temporal forecasting and spatial interpolation was large with the spatial interpolation being far less performant. In the analysis, a possible cause was attributed to the sparsity of the spatial data when compared to the temporal data. That is, the groundwater behaves cyclically and changes relatively slowly over time with respect to the temporal sampling frequency while the changes in groundwater levels are much faster relative to the spatial sampling frequency of the data at hand. If this is the case, a high-impact solution for increasing model performance would be to increase the spatial sampling frequency, and namely, create and service more groundwater wells. This is unfortunately an expensive endeavor which leads to another direction for future work: finding optimal sites for building new groundwater wells.

While the methods outlined in this thesis are one driver for the creation of other wells, there are likely other drivers such as economic cost of building the wells, the need for monitoring in the location driven by factors such as anthropogenic influence and utilization of groundwater in the area. If the goal of the BGR is to minimize the costs while maintaining effective and useful measurements of groundwater levels across Brandenburg, incorporating the performance of operationally utilized forecasting models such as this could be of interest. Therefore, finding optimal sites for new groundwater wells could build on work of the type covered in the thesis.

In addition to finding optimal locations to build more wells, an examination of the performance of the methods on smaller regions with spatially denser data such as the area along the Oder river or in the city of Berlin could be undertaken to provide more context for spatiotemporal modeling. This would also explore the issue of spatial data density and its effect on the spatiotemporal forecasting models.

Finally, additional covariates not yet explored could have a large impact on the model performance. Data about the anthropogenic influences in the area surrounding the monitoring wells such as GDP, irrigation utilization, population, and consumption of water, could be of interest.

6.3.3. Analysis of Errors in Input Data

Another direction for future work would be the analysis of how the models behave when errors are introduced to the data. There has been work on the evaluation of machine learning models with errors in the input data in several recent works [23, 22, 21]. In the problem explored in this thesis, one would need to first develop timeseries specific errors and then evaluate the performance degredation of the models when the model does inference on data with these errors. The data-dependent errors could be implemented with the `tab-err` software [23].

6.3.4. Formalization, Design, and Engineering of Novel Methods

This thesis was primarily scientifically exploratory in nature and in order to have a larger impact on the operational processes in water management in Brandenburg and to ensure that future work can easily compare to the methods tested, some of the newly developed methods should be made deployable. This consists of formalizing the design for the spatiotemporal models in such a way that another scientist or engineer could apply the model in a manner similar to scikit-learn's fit-predict pattern. In order to do this, careful consideration of the data and software dependencies within the modeling must be accounted for and exposed at the correct level of abstraction to ensure the design is extensible while maintaining ease-of-use. Additionally, relevant documentation of the methods must be written to accompany the software, thus contributing to elevated ease-of-use and extensibility for further development.

6.3.5. Further Analysis of Relative Performance in Operational Settings

Closer collaboration with scientists at the German Federal Ministry of Natural Resources in order to better understand the operational challenges such a model could solve are another step to be taken. In specific, an understanding of the software ecosystem in which their current processes exist and how best to contribute the performant developed models to it is necessary. This would require first an evaluation and rigorous comparison of the developed models with the exact processes they utilize and then an implementation of the developed models in the specified environment. Finally, monitoring of the deployment and model drift over time would be necessary to ensure that the developed models continue to improve the performance of the systems used in groundwater management.

Bibliography

[1] Giuseppe Amatulli, Sami Domisch, Mao-Ning Tuanmu, Benoît Parmentier, Ajay Ranipeta, John Malczyk, and Walter Jetz. "A suite of global, cross-scale topographic variables for environmental and biodiversity modeling". In: *Scientific Data* 5.1 (2018). DOI: 10.1038/sdata.2018.40.

[2] Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, et al. "Chronos-2: From Univariate to Universal Forecasting". In: (Oct. 2025). arXiv:2510.15821 [cs.LG]. DOI: 10.48550/arXiv.2510.15821.

[3] Tao Bai and Pejman Tahmasebi. "Graph neural network for groundwater level forecasting". In: *Journal of Hydrology* 616 (2023), p. 128792. DOI: 10.1016/j.jhydrol.2022.128792.

[4] Nico Beck, Jonas Dovern, and Stefanie Vogl. "Mind the naive forecast! A rigorous evaluation of forecasting models for time series with low predictability". In: *Applied Intelligence* 55.6 (2025), p. 395. DOI: 10.1007/s10489-025-06268-w.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 1877–1901.

[6] Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza, Max Mergenthaler-Canseco, and Artur Dubrawski. "N-HiTS: Neural Hierarchical Interpolation for Time Series Forecasting". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 6. arXiv:2201.12886 [cs.LG]. AAAI Press, 2023, pp. 6989–6997. DOI: 10.1609/aaai.v37i6.25854.

[7] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. DOI: 10.3115/v1/d14-1179.

[8] Tim Davie. *Fundamentals of Hydrology*. 2nd. Routledge Fundamentals of Physical Geography. Routledge, 2008. ISBN: 978-0-415-39986-9.

[9] Jan G. De Gooijer and Rob J. Hyndman. “25 years of time series forecasting”. In: *International Journal of Forecasting* 22.3 (2006). Twenty five years of forecasting, pp. 443–473. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2006.01.001.

[10] European Environment Agency. *CORINE Land Cover (CLC) 2018, Version 2020_20u1*. Data set. 2018. URL: https://land.copernicus.eu/pan-european/corine-land-cover.

[11] Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline C Lisaius, Markus Immitzer, David A. Coomes, Anil Madhavapeddy, Andrew Blake, and Srinivasan Keshav. *TESSERA: Temporal Embeddings of Surface Spectra for Earth Representation and Analysis*. 2025. DOI: 10.48550/arXiv.2506.20380. arXiv: 2506.20380 [cs.LG].

[12] Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger, Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. *TabPFN-3: Technical Report*. 2026. DOI: 10.48550/arXiv.2605.13986. arXiv: 2605.13986 [cs.LG].

[13] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. New York, NY, USA: Springer-Verlag, 2009. ISBN: 978-0387848570. URL: https://link.springer.com/book/10.1007/978-0-387-84858-7.

[14] Kevin M. Hiscock and Victor F. Bense. *Hydrogeology: Principles and Practice*. 2nd. Wiley-Blackwell, 2014. ISBN: 978-0-470-65663-1. URL: https://www.wiley.com/en-us/Hydrogeology%3A+Principles+and+Practice%2C+2nd+Edition-p-9780470656631.

[15] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[16] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. “Accurate predictions on small data with a tabular foundation model”. In: *Nature* (Jan. 2025). DOI: 10.1038/s41586-024-08328-6.

[17] Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. “From Tables to Time: Extending TabPFN-v2 to Time Series Forecasting”. In: (Jan. 2026). arXiv:2501.02945 [cs.LG]. DOI: 10.48550/arXiv.2501.02945.

[18] Kurt Hornik. “Approximation capabilities of multilayer feedforward networks”. In: *Neural Networks* 4.2 (1991), pp. 251–257. DOI: 10.1016/0893-6080(91)90009-t.

[19] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "LoRA: Low-Rank Adaptation of Large Language Models". In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[20] Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice (3rd ed)*. Online open-access textbook. OTexts, Apr. 2021. URL: https://otexts.com/fpppy/ (visited on 04/15/2026).

[21] Sebastian Jäger, Arndt Allhorn, and Felix Bießmann. "A Benchmark for Data Imputation Methods". en. In: *Frontiers in Big Data* 4 (July 2021). ISSN: 2624-909X. DOI: 10.3389/fdata.2021.693674.

[22] Philipp Jung, Nicholas Chandler, Sebastian Jäger, and Felix Bießmann. "MechDetect: Detecting Data-Dependent Errors". In: *Proceedings of the 2025 International Conference on Data Science and Intelligent Systems (DSIS)*. IEEE, 2025. DOI: 10.1109/DSIS67228.2025.11390600.

[23] Philipp Jung, Sebastian Jäger, Nicholas Chandler, and Felix Biessmann. "Towards Realistic Error Models for Tabular Data". In: *J. Data and Information Quality* 17.4 (2025). ISSN: 1936-1955. DOI: 10.1145/3774914.

[24] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations (ICLR)*. 2015. DOI: 10.48550/arXiv.1412.6980.

[25] Daniel G. Krige. "A statistical approach to some basic mine valuation problems on the Witwatersrand". In: *Journal of the Chemical, Metallurgical and Mining Society of South Africa* 52.6 (Dec. 1951). Seminal paper in the field of geostatistics., pp. 119–139. URL: https://journals.co.za/doi/pdf/10.10520/AJA0038223X_4792 (visited on 04/15/2026).

[26] Vijay Kumar. "Spatial Interpolation of Groundwater Levels by Kriging". In: *Journal of the Geological Society of India* 69 (May 2007), pp. 996–1004. DOI: 10.17491/jgsi/2007/690514.

[27] Stefan Kunz, Alexander Schulz, Maria Wetzel, Maximilian Nölscher, Teodor Chiaburu, Felix Bießmann, and Stefan Broda. "Towards a global spatial machine learning model for seasonal groundwater level predictions in Germany". en. In: *Hydrology and Earth System Sciences* 29.15 (Aug. 2025), pp. 3405–3433. DOI: 10.5194/hess-29-3405-2025.

[28] Bryan Lim, Sercan Ö. Arik, Nicolas Loeff, and Tomas Pfister. "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting". In: *International Journal of Forecasting* 37.4 (2021). arXiv:1912.09363 [stat.ML], pp. 1748–1764. DOI: 10.1016/j.ijforecast.2021.03.012.

[29] J. K. Lovelace, M. G. Nielsen, A. L. Read, C. J. Murphy, and M. A. Maupin. *Estimated Groundwater Withdrawals from Principal Aquifers in the United States, 2015*. en. Circular 1464. Reston, VA: U.S. Geological Survey, Dec. 2020. DOI: 10.3133/cir1464.

[30] Paula Moraga. *Spatial Statistics for Data Science: Theory and Practice with R*. Chapman & Hall/CRC Data Science Series. ISBN: 9781032633510. Chapman & Hall/CRC, 2023. ISBN: 9781032633510. URL: https://www.paulamoraga.com/book-spatial/ (visited on 04/15/2026).

[31] Pratik Nag, Ying Sun, and Brian J. Reich. "Spatio-temporal DeepKriging for interpolation and probabilistic forecasting". In: *Spatial Statistics* 57 (2023). ISSN: 2211-6753. DOI: 10.1016/j.spasta.2023.100773.

[32] Maximilian Nölscher, Michael Mutz, and Stefan Broda. "Multiorder Hydrologic Position for Europe — a Set of Features for Machine Learning and Analysis in Hydrology". In: *Scientific Data* 9.1 (2022), p. 654. DOI: 10.1038/s41597-022-01787-4.

[33] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. "On the Spectral Bias of Neural Networks". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 5301–5310. DOI: 10.48550/arXiv.1806.08734.

[34] Ali Rahimi and Benjamin Recht. "Random Features for Large-Scale Kernel Machines". In: *Advances in Neural Information Processing Systems*. Ed. by J. Platt, D. Koller, Y. Singer, and S. Roweis. Vol. 20. Curran Associates, Inc., 2007. URL: https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

[35] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. ISBN: 9780262182539. Cambridge, Mass: MIT Press, 2006. ISBN: 9780262182539. URL: http://www.gaussianprocess.org/gpml/ (visited on 04/15/2026).

[36] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, Mohammad Sadegh Salehi, Farzeen Farnia, Habiba Ghonia, Rishikanth Gupta, Shobhit Kalanthadar, Mullen Reddy, et al. "Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting". In: (Feb. 2024). arXiv:2310.08278 [cs.LG]. DOI: 10.48550/arXiv.2310.08278.

[37] Yunus Saatci. "Scalable Inference for Structured Gaussian Process Models". PhD thesis. University of Cambridge, 2011. URL: http://mlg.eng.cam.ac.uk/pub/pdf/Saa11.pdf.

[38] Michelle Sapitang, Wanie M. Ridwan, Ali Najah Ahmed, Chow Ming Fai, and Ahmed El-Shafie. "Groundwater level as an input to monthly predicting of water level using various machine learning algorithms". In: *Earth Science Informatics* 14.3 (Sept. 2021), pp. 1269–1283. DOI: 10.1007/s12145-021-00654-x.

[39] Igor Shiklomanov. "World Fresh Water Resources". en. In: *Water in Crisis: A Guide to the World's Fresh Water Resources*. Ed. by Peter H. Gleick. New York: Oxford University Press, 1993, pp. 13–24. ISBN: 978-0195076271.

[40] Evdokia Tapoglou, George P. Karatzas, Ioannis C. Trichakis, and Emmanouil A. Varouchakis. "A spatio-temporal hybrid neural network-Kriging model for groundwater level simulation". In: *Journal of Hydrology* 519 (2014), pp. 3193–3203. DOI: 10.1016/j.jhydrol.2014.10.040.

[41] Waldo R. Tobler. "A Computer Movie Simulating Urban Growth in the Detroit Region". In: *Economic Geography* 46 (June 1970). Supplement: Proceedings. International Geographical Union. Commission on Quantitative Methods, pp. 234–240. DOI: 10.2307/143141.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. arXiv:1706.03762 [cs]. Curran Associates, Inc., 2017. DOI: 10.48550/arXiv.1706.03762.

[43] William W. Woessner and Eileen P. Poeter. *Hydrogeologic Properties of Earth Materials and Principles of Groundwater Flow*. Guelph, Ontario, Canada: The Groundwater Project, 2020. ISBN: 978-1-7770541-2-0. DOI: 10.21083/978-1-7770541-2-0. URL: https://books.gw-project.org/hydrogeologic-properties-of-earth-materials-and-principles-of-groundwater-flow/ (visited on 06/22/2026).

[44] Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. *Dive into Deep Learning*. Online interactive book. Cambridge University Press, 2023. URL: https://d2l.ai/ (visited on 04/15/2026).

[45] Yanxin Zhu and Klaus-Dieter Balke. "Groundwater Protection: What Can We Learn from Germany?" en. In: *Journal of Zhejiang University Science B* 9.3 (Mar. 2008), pp. 227–231. DOI: 10.1631/jzus.B0710639.

A. Temporal Forecasting on German Multi-State Dataset

The dataset considered here is one in which the entirety of Germany (except Saarland) is covered and the temporal history goes from 1990 to 2016 with weekly measurements. It contains covariates describing the geological region surrounding the measurement wells as well as the environmental variables that vary with time such as temperature. Temporal modeling and deeper analysis of the dataset is done in [27] where additional details may be found.

A.1. Dataset Description

In this section, details from Kunz et al. [27] are reviewed and additional analyses are conducted to put the rest of the work in better context. It begins with explaining how the aforementioned paper did temporal data splits, then provides visualizations of the spatial distribution of the wells and density of the time series data. Then an individual time series of the target variable, groundwater level, is given from a selected well.

When training any predictive or forecasting model, it is important to create subsets of the data so generalization performance of the models can be effectively understood. To this end, in [27] the authors split the data temporally with a training set ranging from 1990 to 2010, a validation set ranging from 2010 to 2013, and a test set ranging from 2013 to 2016. A visualization worthy of investigation is that of the temporal forecast target, the groundwater level. Below, in Figure A.1 is an example of the historical groundwater level (m NHN) from the training set at a site in Brandenburg with ID: BB-25470023.

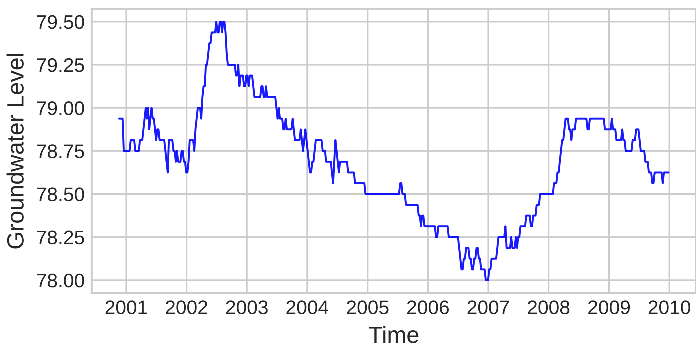

Figure A.1.: Historical Groundwater Level at Site BB-25470023

The covariates in the dataset are explored in more detail in [27]. It is noteworthy that each of the series in the given training set have varying histories. We investigate the histories and spatial distribution of the groundwater level time series in Figure A.2. This visualization depicts the length of the time series from the training, validation, and test sets, concatenated where possible and standalone where not. The key insight is that there are some regions with density different from other regions. Particular attention should be drawn to the region of Thuringia where each timeseries has approximately the same length which is below the average.

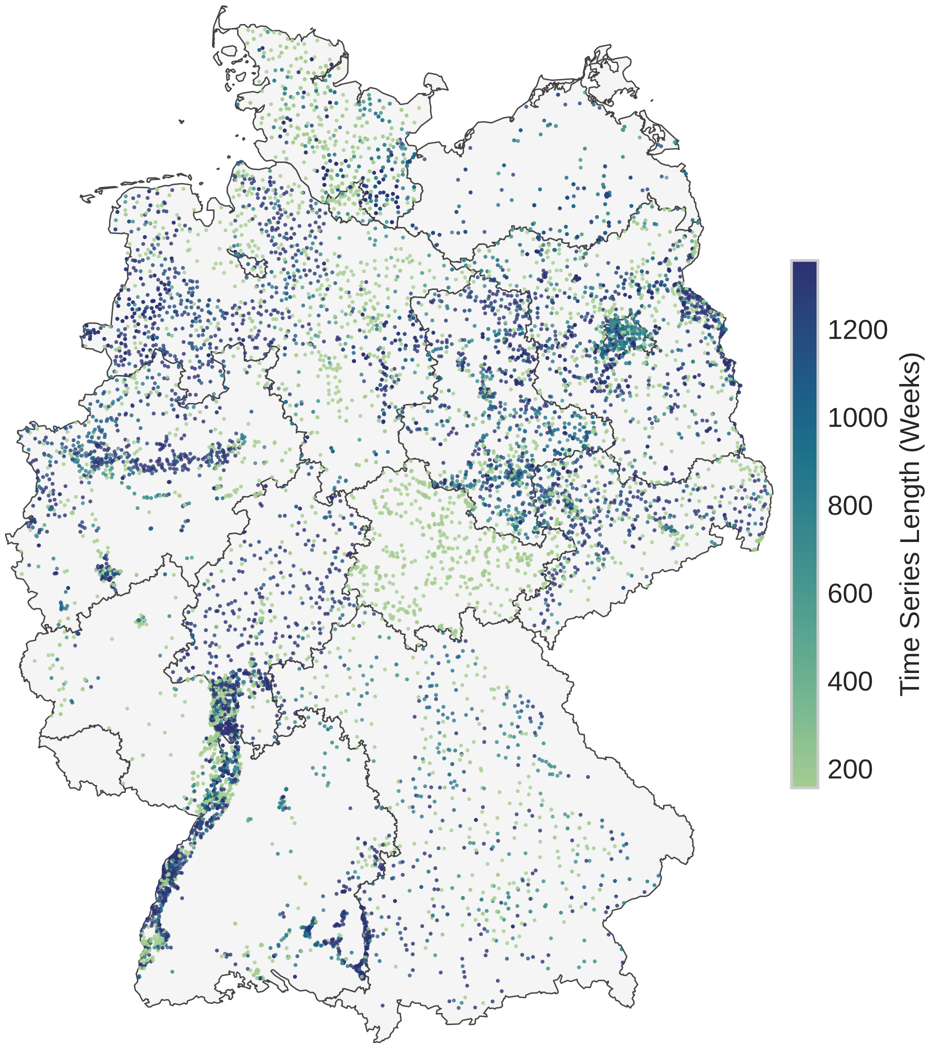

Figure A.2.: Groundwater Level Timeseries Length by Location in Germany

In this section, we provided some analysis on a secondary dataset that has some spatial and temporal overlap with the dataset chosen. The data were given with splits already made as in [27]. A sample groundwater level timeseries from the monitoring well BB-25470023 was visualized as was the spatial distribution of the data throughout Germany.

A.2. Temporal Modeling

A subset of the results from [27] were replicated using the root mean squared error and temporal fusion transformer described therein. The problem proposed in [27] was to create a global forecasting model to forecast groundwater levels in horizons of one and twleve weeks throughout Germany. To accomplish this, they utilized temporal fusion transformers and N-HiTS. In our experiments, we replicate the setup in that we primarily use a context of 52 weeks and forecast the same horizon of twelve weeks. The results here are meant to verify those found in the paper and we refer the reader to the paper for further information regarding the other forecast horizons, models, and metrics. We also utilize the same covariates as those given in the paper. In order to evaluate the error of the forecasting model relative to the true futures, we utilize the root mean squared error on windowed predictions with a stride of one. The median of the root mean squared error across all wells for which forecasts were made is the final reported metric for each of the models evaluated. Specifically, the error is evaluated on the validation set from the paper, the time period from the beginning of 2010 to the end of 2012. In order to ensure the predictive models had sufficient context, the last 52 weeks of the training set were prepended to the validation set but forecasts were not made on this time period. We now discuss some additional characteristics of the data that are relevant to the forecasting.

The temporal fusion transformer as in [27] was trained with the hyperparameters in the following table A.1.

Table A.1.: Hyperparameter Configuration for Temproal Fusion Transformer

Hyperparameter	**Value**	**Description**
Learning Rate (η)	0.003981	Step size for weight updates
Batch Size	8192	Number of samples per iteration
Learning Rate Scheduler	SWA	Stochastic Weight Averaging
Context Length	52	Number of past observations (lookback)
Prediction Horizon	12	Number of future steps to forecast
Gradient Clip Value	0.2	Value to clip gradients at
Early Stopping Delta	0.0001	Minimum loss change tolerance
Early Stopping Patience	5	Maximum iterations before stopping
Loss Function	Quantile (0.5)	Guide for learning optimization

The training took about 12 hours on an H200 Nvidia GPU the inference also took 12 hours. The median root mean squared error, for the forecast horizon of twelve windowed on the validation set is: 0.156. This is approximately what the paper detailed in the results section and serves as a baseline for the subsequent experiments. In the remainder of this section, the model utilized is Chronos2.

The details of the Chronos2 model are covered in more depth in the related work section. The most pertinent information about it for our purposes is that it is a time series foundation model capable of multi-series, covariate-informed zero shot as well as fine tuned forecasting. It is trained with a quantile loss but is used here to forecast the median value (0.5th quantile) of the time series.

The context length was 52 weeks for this experiment and the horizons of twelve and one weeks were used. The cross-learning function of Chronos2 was also applied which allows for sharing of information between the time series. The resulting median root mean squared error, for the forecast horizon of twelve weeks windowed on the validation set was approximately: 0.217. The resulting median root mean squared error, for the corresponding forecast horizon of 1 windowed on the validation set was approximately: 0.0815. Inference on the entire validation set These results show worse performance than the temporal fusion transformer which is somewhat expected due to the lack of specific training for this problem.

In order to perform finetuning of the Chronos2 model, one must first do hyperparameter optimzation. The search space is detailed in the following table A.2.

Table A.2.: Hyperparameter Search Space for Chronos2

Hyperparameter	**Space**	**Description**
Learning Rate (η)	$\left[10^{-6}, 10^{-4}\right]$	Step size for weight updates
Number of Steps	$[500, 2000]$	Number of steps to fine tune
Batch Size	$\{32, 64, 128\}$	Number of samples per iteration
LoRA	True/False	Use Low Rank Adaptation
LoRA α	{4, 8, 16, 32}	Scaling of adapter weights
LoRA r	{16, 32, 64}	Rank of adapter, contols expressiveness

Both the horizon twelve and horizon one hyperparameter optimization runs had 15 trials each and used the bayesian optimization strategy. The resulting hyperparameters are given in tables A.3 and A.4. Both approaches optimized the quantile loss and had a fixed context length of 52.

Table A.3.: Horizon 1 Hyperparameter Selection for Chronos2

Hyperparameter	**Value**
Learning Rate (η)	8×10^{-5}
Number of Steps	1100
Batch Size	64
LoRA	False
LoRA α	N/A
LoRA r	N/A

After fine tuning using the hyperparameters found during hyperparameter optimization, the median root mean squared error across all wells and all windowed forecasts is approximately: 0.072 and 0.1495 for the horizons of one week and twelve weeks respectively. This yields slight outperformance of the temporal fusion transformer architecture in the twelve week horizon scenario and underperformance (with reference to [27]) in the one week horizon scenario where the temporal fuison transformer obtained a median performance

Table A.4.: Horizon 12 Hyperparameter Selection for Chronos2

Hyperparameter	**Value**
Learning Rate (η)	7.6×10^{-5}
Number of Steps	2000
Batch Size	128
LoRA	False
LoRA α	N/A
LoRA r	N/A

of 0.06 root mean squared error. When compared to the N-HiTS model from the paper, Chronos2 underperforms in both horizons.

A final experiment that was run on this data was the investigation of the effect that the context length had on the zero shot performance of Chronos2. In particular, the context length was altered, ranging between: {4, 12, 26, 52} weeks. The other parameters such as the horizon, and the cross learning were kept the same. The results in median root mean squared error are summarized in the following table A.5.

Table A.5.: Zero Shot Data Efficiency Results

	Context			
Horizon	4	12	26	52
1	0.0764	0.0818	0.0866	0.0814
12	0.221	0.225	0.237	0.217

There appears to be a descent in performance until the context length of 52 weeks is reached in both forecast horizons. This could be due to the model “getting confused” with more data and then, when it has a long enough context window, being able to utilize a pattern that was not present in the shorter context inputs. This could serve as motivation for more explainability research in the direction of time series foundation models.

In the preceeding sections, we replicated a subset of the results of [27] and evaluated the performance of the Chronos2 time series foundation model on the same task. The zero shot capabilities of the Chronos2 model underperformed the N-HiTS and temporal fusion transformer models used in the paper on both forecast horizons. After fine tuning, the Chronos2 model outperformed the temporal fusion transformer model in the twelve week forecast horizon and under performed in the one week horizon. The N-HiTS model from the paper remains the best-performer in this task. Additionally, the data efficiency of the Chronos2 zero shot capability was assessed through varying the context length and measuring the forecast performance. The data efficiency results appeared to be unintuitive with decreasing performance as the context window was increased until an increase at the 52 week mark.

Due to a change of direction and the realization that smaller spatial areas would be more easily modeled, only temporal forecasting was conducted on the data used in this section before the decision was made to use an alternative dataset that was better suited to meet the goals of the thesis. By investigating the dataset in this section, we verified the results of prior work, and examined the use of novel temporal models on the same task. Additionally, these analyses provide additional context to those in the thesis.

B. Data

In this chapter we present additional analyses and figures deemed less vital to the main text.

B.1. GeoTessera PCA

More details on the GeoTessera embeddings and the corresponding PCA are given here. The Cumulative Proportion of Variance Explained plot is given in Figure B.1.

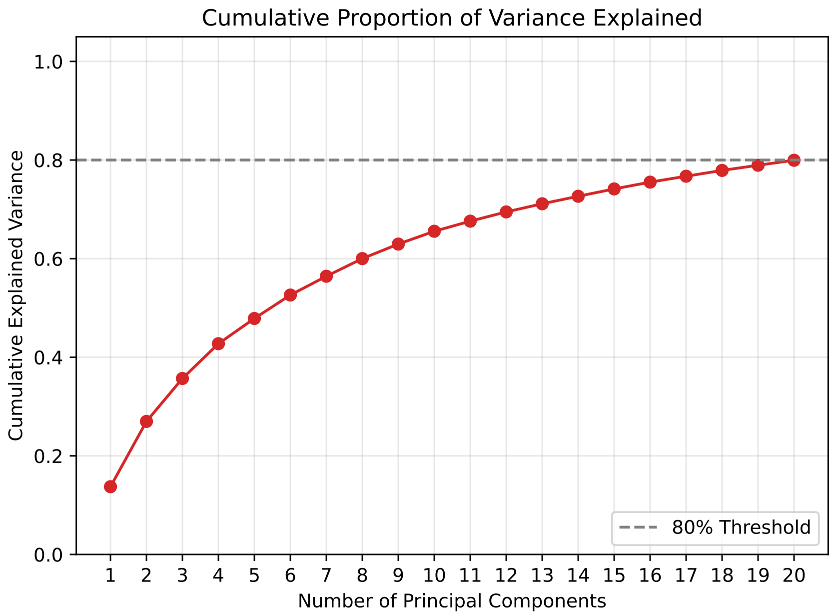

Figure B.1.: Cumulative Proportion of Variance Explained of GeoTessera Embeddings PCA

The biplot associated with the first two principal components is given in Figure B.2

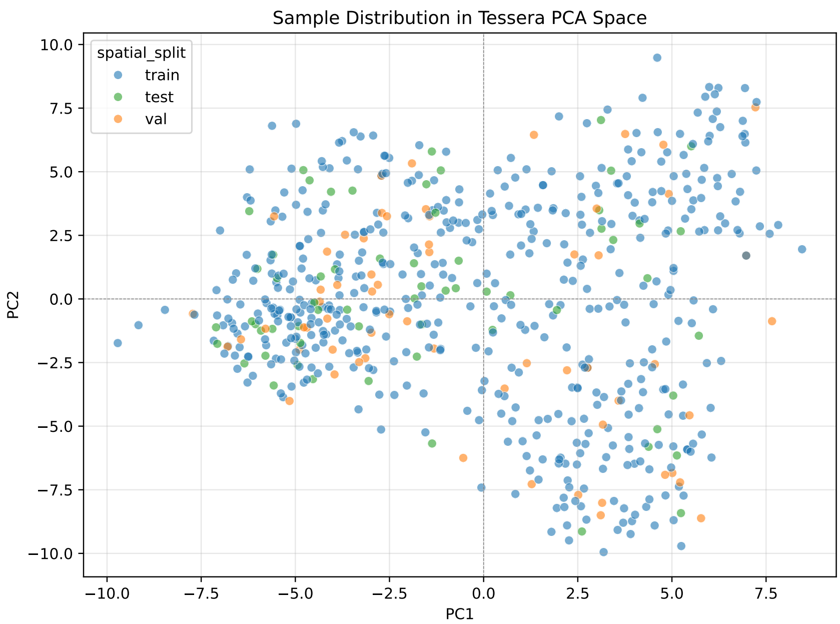

Figure B.2.: Biplot of GeoTessera Embeddings PCA. Principal Components 1 and 2.

C. Mathematical Appendicies

This chapter contains any relevant derivations or mathematics relevant to the main text.

C.1. Kronecker Gaussian Process Derivation

A more detailed derivation of the Kronecker Gaussian Process is given here.
Component Eigendecomposition Mechanics:

Because the independent spatial covariance $\mathbf{K}_s \in \mathbb{R}^{N\times N}$ and temporal covariance $\mathbf{K}_t \in \mathbb{R}^{T\times T}$ are real and symmetric, the spectral theorem guarantees their orthogonal diagonalization:

$$\begin{aligned}\mathbf{K}_s &= \mathbf{Q}_s\mathbf{\Lambda}_s\mathbf{Q}_s^\top \quad \text{subject to} \quad \mathbf{Q}_s^\top\mathbf{Q}_s = \mathbf{I}_N,\\ \mathbf{K}_t &= \mathbf{Q}_t\mathbf{\Lambda}_t\mathbf{Q}_t^\top \quad \text{subject to} \quad \mathbf{Q}_t^\top\mathbf{Q}_t = \mathbf{I}_T\end{aligned}$$

where $\mathbf{Q}_s, \mathbf{Q}_t$ are the orthonormal eigenvector matrices, and $\boldsymbol{\lambda}_s = \text{diag}(\mathbf{\Lambda}_s)$, $\boldsymbol{\lambda}_t = \text{diag}(\mathbf{\Lambda}_t)$ are the isolated eigenvalue vectors.
Joint Eigenspace and Identity Mapping:

By leveraging the distributive property of the Kronecker product over standard matrix multiplication, $(\mathbf{AB})\otimes(\mathbf{CD}) = (\mathbf{A}\otimes\mathbf{C})(\mathbf{B}\otimes\mathbf{D})$, the global clean training covariance matrix factors as:

$$\begin{aligned}\mathbf{K}_{cc} &= \mathbf{K}_s \otimes \mathbf{K}_t\\ &= (\mathbf{Q}_s\mathbf{\Lambda}_s\mathbf{Q}_s^\top) \otimes (\mathbf{Q}_t\mathbf{\Lambda}_t\mathbf{Q}_t^\top)\\ &= (\mathbf{Q}_s \otimes \mathbf{Q}_t)(\mathbf{\Lambda}_s \otimes \mathbf{\Lambda}_t)(\mathbf{Q}_s^\top \otimes \mathbf{Q}_t^\top)\end{aligned}$$

Applying the transposition property $(\mathbf{A}\otimes\mathbf{B})^\top = \mathbf{A}^\top \otimes \mathbf{B}^\top$, the joint eigenvector matrix $\mathbf{Q} = \mathbf{Q}_s \otimes \mathbf{Q}_t$ preserves strict orthogonality in the higher-dimensional space:

$$\begin{aligned}\mathbf{Q}^\top\mathbf{Q} &= (\mathbf{Q}_s \otimes \mathbf{Q}_t)^\top(\mathbf{Q}_s \otimes \mathbf{Q}_t)\\ &= (\mathbf{Q}_s^\top\mathbf{Q}_s) \otimes (\mathbf{Q}_t^\top\mathbf{Q}_t)\\ &= \mathbf{I}_N \otimes \mathbf{I}_T\\ &= \mathbf{I}_{NT}\end{aligned}$$

Factoring the homoscedastic observation noise $\sigma_n^2\mathbf{I}_{NT}$ into the core diagonalized system yields:

$$\begin{aligned}\mathbf{K}_{cc} + \sigma_n^2\mathbf{I}_{NT} &= (\mathbf{Q}_s \otimes \mathbf{Q}_t)(\mathbf{\Lambda}_s \otimes \mathbf{\Lambda}_t)(\mathbf{Q}_s \otimes \mathbf{Q}_t)^\top + \sigma_n^2(\mathbf{Q}_s \otimes \mathbf{Q}_t)\mathbf{I}_{NT}(\mathbf{Q}_s \otimes \mathbf{Q}_t)^\top\\ &= (\mathbf{Q}_s \otimes \mathbf{Q}_t)\left[\mathbf{\Lambda}_s \otimes \mathbf{\Lambda}_t + \sigma_n^2\mathbf{I}_{NT}\right](\mathbf{Q}_s \otimes \mathbf{Q}_t)^\top\end{aligned}$$

where the internal diagonal entries correspond to the row-major vectorization of the compact 2D grid matrix $\mathbf{L}_{\text{noisy}} = \boldsymbol{\lambda}_s\boldsymbol{\lambda}_t^\top + \sigma_n^2\mathbf{1}\mathbf{1}^\top$.

Derivation of the Eigenspace Target Rotation:

To evaluate the rotated target vector $\mathbf{y}_{\text{rot}} = \mathbf{Q}^\top \mathbf{y}$, the row-major vectorization identity $(\mathbf{A} \otimes \mathbf{B})\text{vec}_R(\mathbf{X}) = \text{vec}_R(\mathbf{A}\mathbf{X}\mathbf{B}^\top)$ is applied to map the operation from $\mathbb{R}^{NT}$ into $\mathbb{R}^{N\times T}$ matrix space:

$$\begin{aligned}
\mathbf{y}_{\text{rot}} &= (\mathbf{Q}_s \otimes \mathbf{Q}_t)^\top \text{vec}_R(\mathbf{Y}) \\
&= (\mathbf{Q}_s^\top \otimes \mathbf{Q}_t^\top)\text{vec}_R(\mathbf{Y}) \\
&= \text{vec}_R\left(\mathbf{Q}_s^\top \mathbf{Y}(\mathbf{Q}_t^\top)^\top\right) \\
&= \text{vec}_R\left(\mathbf{Q}_s^\top \mathbf{Y}\mathbf{Q}_t\right)
\end{aligned}$$

Isolating the underlying 2D tensor matches the matrix-space target rotation configuration:

$$\mathbf{Y}_{\text{rot}} = \mathbf{Q}_s^\top \mathbf{Y}\mathbf{Q}_t$$

System Inversion and Back-Projection:

Substituting the spectral components back into the full linear weight system allows the execution of the matrix inversion step exclusively along the diagonalized eigenspace:

$$\begin{aligned}
\boldsymbol{\alpha} &= (\mathbf{K}_{cc} + \sigma_n^2 \mathbf{I}_{NT})^{-1}\mathbf{y} \\
&= (\mathbf{Q}_s \otimes \mathbf{Q}_t)\left[\boldsymbol{\Lambda}_s \otimes \boldsymbol{\Lambda}_t + \sigma_n^2 \mathbf{I}_{NT}\right]^{-1}(\mathbf{Q}_s \otimes \mathbf{Q}_t)^\top \text{vec}_R(\mathbf{Y}) \\
&= (\mathbf{Q}_s \otimes \mathbf{Q}_t)\left[\boldsymbol{\Lambda}_s \otimes \boldsymbol{\Lambda}_t + \sigma_n^2 \mathbf{I}_{NT}\right]^{-1}\text{vec}_R(\mathbf{Y}_{\text{rot}})
\end{aligned}$$

Because the inverted core matrix is strictly diagonal, this operation simplifies to an element-wise Hadamard division ($| \oslash |$) by the 2D joint eigenvalue grid:

$$\begin{aligned}
\boldsymbol{\alpha} &= (\mathbf{Q}_s \otimes \mathbf{Q}_t)\text{vec}_R\left(\mathbf{Y}_{\text{rot}} \oslash \mathbf{L}_{\text{noisy}}\right) \\
&= \text{vec}_R\left(\mathbf{Q}_s\left[\mathbf{Y}_{\text{rot}} \oslash \mathbf{L}_{\text{noisy}}\right]\mathbf{Q}_t^\top\right)
\end{aligned}$$

Resolving the final row-major identity yields the exact target weights array $\mathbf{A} = \mathbf{Q}_s\left(\mathbf{Y}_{\text{rot}} \oslash \mathbf{L}_{\text{noisy}}\right)\mathbf{Q}_t^\top$, concluding the derivation.

D. Additional Results

This section contains additional results including tables of metrics and hyperparameters on validation and test sets.

D.1. Temporal

This section contains additional results for the temporal forecasting task.

Table D.1.: Performance Metrics for Temporal Test Set (Horizon 1) — Median [IQR]

Model	RMSE	NSE	KGE	MPIW	Winkler	PICP
Baseline	0.034 [0.031]	0.965 [0.032]	0.969 [0.026]	—	—	—
C2-ZS	0.035 [0.034]	0.962 [0.038]	0.931 [0.029]	0.388 [0.320]	0.392 [0.330]	1.000 [0.000]
Cs2-FT	0.030 [0.025]	0.971 [0.030]	0.962 [0.029]	0.233 [0.194]	0.243 [0.201]	1.000 [0.014]
N-HITS	0.025 [0.023]	0.980 [0.027]	0.979 [0.022]	0.169 [0.126]	0.212 [0.191]	0.973 [0.041]

Table D.2.: Performance Metrics for Temporal Test Set (Horizon 12) — Median [IQR]

Model	RMSE	NSE	KGE	MPIW	Winkler	PICP
Baseline	0.235 [0.182]	-0.627 [0.763]	0.191 [0.300]	—	—	—
C2-ZS	0.209 [0.169]	-0.290 [0.798]	0.427 [0.306]	1.057 [0.846]	1.515 [1.414]	0.959 [0.095]
C2-FT	0.123 [0.091]	0.555 [0.336]	0.727 [0.198]	0.657 [0.534]	0.808 [0.583]	1.000 [0.054]
N-HITS	0.112 [0.075]	0.655 [0.365]	0.740 [0.214]	0.595 [0.408]	0.712 [0.588]	0.986 [0.054]

D.2. Decoupled Spatiotemporal

This section contains additional results on the decoupled spatiotemporal modeling.

D.2.1. Gaussian Processes on Coordinates

Additional hyperparameter summaries and test metrics for the Gaussian processes on coordinates are contained here.

Table D.3.: Hyperparameters for Coordinate-Only GP on Test Set (Horizon 1)

Hyperparameters	Mean	Std	Median	Q25	Q75
Noise Variance σ_n^2 (H1)	0.0170	0.0001	0.0170	0.0169	0.0172
Spatial Lengthscale (averaged) ℓ	0.2655	0.0006	0.2655	0.2650	0.2660

Table D.4.: Hyperparameters for Coordinate-Only GP on Test Set (Horizon 12)

Hyperparameters	Mean	Std	Median	Q25	Q75
Noise Variance σ_n^2 (H12)	0.0170	0.0002	0.0171	0.0169	0.0172
Spatial Lengthscale (averaged) ℓ	0.2651	0.0008	0.2650	0.2645	0.2657

Table D.5.: Test Set Metrics for Coordinates Only GP

	Metric	RMSE		NSE		KGE	
Model	Horizon	Median	IQR	Median	IQR	Median	IQR
Baseline	H1	1.024	2.142	-32.102	238.125	0.535	0.618
	H12	0.976	2.159	-33.714	197.216	0.056	0.645
N-HiTS	H1	1.023	2.137	-31.630	238.116	0.526	0.618
	H12	0.987	2.166	-33.507	201.818	0.435	0.504
C2 ZS	H1	1.021	2.140	-32.338	237.667	0.516	0.679
	H12	1.022	2.149	-34.832	198.924	0.233	0.670
C2 FT	H1	1.032	2.141	-33.703	236.746	0.508	0.666
	H12	1.002	2.141	-32.060	207.528	0.425	0.624

	Metric	MPIW		PICP		Winkler	
Model	Horizon	Median	IQR	Median	IQR	Median	IQR
Baseline	H1	20.482	10.127	1.000	0.000	20.768	11.616
	H12	20.484	10.182	1.000	0.000	20.778	11.691
N-HiTS	H1	20.499	10.102	1.000	0.000	20.779	11.588
	H12	20.490	10.139	1.000	0.000	20.779	11.625
C2 ZS	H1	20.478	10.111	1.000	0.000	20.761	11.592
	H12	20.495	10.166	1.000	0.000	20.786	11.670
C2 FT	H1	20.528	10.218	1.000	0.000	20.826	11.729
	H12	20.491	10.124	1.000	0.000	20.778	11.610

D.2.2. Gaussian Processes on Coordinates and Geologic Principal Components

Additional hyperparameter summaries and test metrics for the Gaussian processes on coordinates and geologic principal components are contained here.

Table D.6.: Hyperparameters for Coords + Geologic PCs GP on Test Set (Horizon 1)

Hyperparameters	Mean	Std	Median	Q25	Q75
Noise Variance σ_n^2 (H1)	0.0011	0.0001	0.0011	0.0010	0.0012
Spatial Lengthscale ℓ_x	0.5581	0.0545	0.5410	0.5106	0.6131
Spatial Lengthscale ℓ_y	0.5094	0.0118	0.5125	0.5030	0.5186
Geologic Lengthscale ℓ_{PC1}	2.6421	0.0737	2.6631	2.5627	2.7032
Geologic Lengthscale ℓ_{PC2}	6.0048	0.5505	6.2418	5.2760	6.5164
Geologic Lengthscale ℓ_{PC3}	4.4671	0.4371	4.6575	3.8827	4.8428
Geologic Lengthscale ℓ_{PC4}	7.2023	0.7164	7.4782	6.2537	7.8826
Geologic Lengthscale ℓ_{PC5}	3.0354	0.1045	3.0746	2.9236	3.1102
Geologic Lengthscale ℓ_{PC6}	6.2046	0.6555	6.4987	5.3582	6.7615
Geologic Lengthscale ℓ_{PC7}	5.4720	0.6526	5.8168	4.5999	5.9758
Geologic Lengthscale ℓ_{PC8}	1.9115	0.0999	1.9149	1.7996	2.0129

Table D.7.: Hyperparameters for Coords + Geologic PCs GP on Test Set (Horizon 12)

Hyperparameters	Mean	Std	Median	Q25	Q75
Noise Variance σ_n^2 (H12)	0.0010	0.0001	0.0010	0.0009	0.0011
Spatial Lengthscale ℓ_x	0.5219	0.0292	0.5151	0.5076	0.5329
Spatial Lengthscale ℓ_y	0.4811	0.0195	0.4788	0.4653	0.4932
Geologic Lengthscale ℓ_{PC1}	2.4704	0.1502	2.5343	2.3111	2.5855
Geologic Lengthscale ℓ_{PC2}	6.5039	0.1216	6.5048	6.3848	6.6128
Geologic Lengthscale ℓ_{PC3}	4.9476	0.1681	4.9072	4.8060	5.0908
Geologic Lengthscale ℓ_{PC4}	8.1146	0.3629	8.0226	7.7435	8.5704
Geologic Lengthscale ℓ_{PC5}	2.6904	0.2259	2.7864	2.4433	2.8628
Geologic Lengthscale ℓ_{PC6}	6.9232	0.2384	6.8757	6.7020	7.1691
Geologic Lengthscale ℓ_{PC7}	6.3576	0.2986	6.2886	6.0979	6.6612
Geologic Lengthscale ℓ_{PC8}	1.7943	0.0568	1.8080	1.7382	1.8454

Table D.8.: Test Set Metrics for Coords + Geologic PCs GP

	Metric	RMSE		NSE		KGE	
Model	Horizon	Median	IQR	Median	IQR	Median	IQR
Baseline	H1	1.368	2.006	-57.809	172.326	0.475	0.788
	H12	1.343	1.795	-45.067	158.348	-0.030	0.635
N-HiTS	H1	1.380	1.996	-55.167	168.292	0.420	0.931
	H12	1.325	1.854	-44.406	140.734	0.359	0.577
C2 ZS	H1	1.284	2.308	-47.353	179.546	0.490	0.929
	H12	1.269	1.757	-60.678	210.148	0.061	0.638
C2 FT	H1	1.260	1.981	-40.240	148.713	0.396	1.034
	H12	1.331	1.803	-48.713	157.621	0.351	0.573

	Metric	MPIW		PICP		Winkler	
Model	Horizon	Median	IQR	Median	IQR	Median	IQR
Baseline	H1	10.908	6.297	1.000	0.000	11.639	7.736
	H12	11.131	6.506	1.000	0.000	11.967	7.845
N-HiTS	H1	10.972	6.403	1.000	0.000	11.745	7.827
	H12	11.119	6.473	1.000	0.000	11.891	7.969
C2 ZS	H1	11.564	7.546	1.000	0.000	13.713	7.684
	H12	11.381	6.626	1.000	0.000	12.209	7.745
C2 FT	H1	11.128	6.425	1.000	0.000	12.035	7.876
	H12	11.245	6.513	1.000	0.000	11.787	7.885

D.2.3. Gaussian Processes on Coordinates, Geologic Principal Components, and Tessera Principal Components

Additional hyperparameter summaries and test metrics for the Gaussian processes on coordinates, geologic principal components, and GeoTessera principal components are contained here.

Table D.9.: Hyperparameters for Coords + Geologic PCs + Tessera PCs GP on Test Set (Horizon 1)

Hyperparameters	Mean	Std	Median	Q25	Q75
Noise Variance σ_n^2 (H1)	0.0020	0.0008	0.0019	0.0013	0.0023
Spatial Lengthscale ℓ_x	0.5315	0.0313	0.5320	0.5060	0.5549
Spatial Lengthscale ℓ_y	0.4449	0.0291	0.4495	0.4162	0.4664
Geologic Lengthscale ℓ_{PC1}	2.4487	0.1847	2.5141	2.2047	2.6272
Geologic Lengthscale ℓ_{PC2}	6.5417	0.0368	6.5372	6.5124	6.5649
Geologic Lengthscale ℓ_{PC3}	4.3966	0.0431	4.4045	4.3612	4.4301
Geologic Lengthscale ℓ_{PC4}	7.6215	0.0772	7.6304	7.5278	7.7051
Geologic Lengthscale ℓ_{PC5}	2.8291	0.0640	2.8573	2.7470	2.8817
Geologic Lengthscale ℓ_{PC6}	8.1203	0.0780	8.1231	8.0614	8.1768
Geologic Lengthscale ℓ_{PC7}	9.7774	0.1210	9.7912	9.6331	9.9083
Geologic Lengthscale ℓ_{PC8}	2.6231	0.0299	2.6226	2.5992	2.6428
Tessera Lengthscale ℓ_{T0}	5.0153	0.0662	5.0432	4.9364	5.0681
Tessera Lengthscale ℓ_{T1}	4.8794	0.0636	4.9102	4.7981	4.9293
Tessera Lengthscale ℓ_{T2}	4.8125	0.0597	4.8347	4.7493	4.8591
Tessera Lengthscale ℓ_{T3}	5.5196	0.0764	5.5496	5.4268	5.5822
Tessera Lengthscale ℓ_{T4}	5.0299	0.0677	5.0588	4.9470	5.0823
Tessera Lengthscale ℓ_{T5}	5.1682	0.0823	5.1926	5.0853	5.2299
Tessera Lengthscale ℓ_{T6}	4.5973	0.0605	4.6156	4.5433	4.6451
Tessera Lengthscale ℓ_{T7}	5.4779	0.0683	5.4996	5.4038	5.5295
Tessera Lengthscale ℓ_{T8}	4.5923	0.0479	4.5927	4.5579	4.6251
Tessera Lengthscale ℓ_{T9}	4.9237	0.0591	4.9498	4.8519	4.9723
Tessera Lengthscale ℓ_{T10}	4.6649	0.0654	4.6868	4.5917	4.7145
Tessera Lengthscale ℓ_{T11}	5.1792	0.0682	5.1977	5.1158	5.2326
Tessera Lengthscale ℓ_{T12}	5.4389	0.0803	5.4748	5.3386	5.5037
Tessera Lengthscale ℓ_{T13}	5.4742	0.0753	5.5026	5.3759	5.5358
Tessera Lengthscale ℓ_{T14}	6.2122	0.0952	6.2525	6.0857	6.2996
Tessera Lengthscale ℓ_{T15}	4.7457	0.0566	4.7637	4.6827	4.7872
Tessera Lengthscale ℓ_{T16}	5.7943	0.0810	5.8269	5.6885	5.8678
Tessera Lengthscale ℓ_{T17}	5.6170	0.0786	5.6499	5.5162	5.6859
Tessera Lengthscale ℓ_{T18}	4.2726	0.0657	4.2935	4.2037	4.3202
Tessera Lengthscale ℓ_{T19}	5.3612	0.0755	5.3913	5.2688	5.4215

Table D.10.: Hyperparameters for Coords + Geologic PCs + Tessera PCs GP on Test Set (Horizon 12)

Hyperparameters	Mean	Std	Median	Q25	Q75
Noise Variance σ_n^2 (H12)	0.0016	0.0008	0.0013	0.0011	0.0020
Spatial Lengthscale ℓ_x	0.5612	0.0390	0.5545	0.5409	0.5695
Spatial Lengthscale ℓ_y	0.4710	0.0373	0.4632	0.4534	0.4751
Geologic Lengthscale ℓ_{PC1}	2.2662	0.0798	2.2201	2.2114	2.3468
Geologic Lengthscale ℓ_{PC2}	6.5031	0.0302	6.5052	6.4843	6.5256
Geologic Lengthscale ℓ_{PC3}	4.3490	0.0253	4.3451	4.3288	4.3683
Geologic Lengthscale ℓ_{PC4}	7.5219	0.0217	7.5192	7.5032	7.5393
Geologic Lengthscale ℓ_{PC5}	2.7585	0.0214	2.7524	2.7431	2.7716
Geologic Lengthscale ℓ_{PC6}	8.0192	0.0374	8.0223	7.9894	8.0465
Geologic Lengthscale ℓ_{PC7}	9.6166	0.0336	9.6140	9.5890	9.6482
Geologic Lengthscale ℓ_{PC8}	2.6068	0.0298	2.5981	2.5859	2.6219
Tessera Lengthscale ℓ_{T0}	4.9370	0.0194	4.9373	4.9236	4.9511
Tessera Lengthscale ℓ_{T1}	4.7992	0.0168	4.7999	4.7888	4.8131
Tessera Lengthscale ℓ_{T2}	4.7372	0.0217	4.7396	4.7233	4.7532
Tessera Lengthscale ℓ_{T3}	5.4287	0.0192	5.4287	5.4152	5.4393
Tessera Lengthscale ℓ_{T4}	4.9438	0.0205	4.9422	4.9296	4.9552
Tessera Lengthscale ℓ_{T5}	5.0666	0.0399	5.0704	5.0360	5.1013
Tessera Lengthscale ℓ_{T6}	4.5179	0.0265	4.5210	4.5018	4.5404
Tessera Lengthscale ℓ_{T7}	5.3878	0.0263	5.3882	5.3696	5.4040
Tessera Lengthscale ℓ_{T8}	4.5520	0.0299	4.5544	4.5306	4.5707
Tessera Lengthscale ℓ_{T9}	4.8487	0.0183	4.8530	4.8370	4.8622
Tessera Lengthscale ℓ_{T10}	4.5860	0.0255	4.5864	4.5712	4.5976
Tessera Lengthscale ℓ_{T11}	5.0912	0.0346	5.0983	5.0747	5.1159
Tessera Lengthscale ℓ_{T12}	5.3373	0.0211	5.3384	5.3242	5.3532
Tessera Lengthscale ℓ_{T13}	5.3782	0.0224	5.3767	5.3627	5.3964
Tessera Lengthscale ℓ_{T14}	6.0949	0.0209	6.1009	6.0771	6.1119
Tessera Lengthscale ℓ_{T15}	4.6737	0.0213	4.6741	4.6606	4.6871
Tessera Lengthscale ℓ_{T16}	5.6944	0.0182	5.6981	5.6776	5.7076
Tessera Lengthscale ℓ_{T17}	5.5160	0.0211	5.5192	5.4973	5.5316
Tessera Lengthscale ℓ_{T18}	4.1968	0.0288	4.1968	4.1814	4.2123
Tessera Lengthscale ℓ_{T19}	5.2615	0.0267	5.2620	5.2411	5.2807

Table D.11.: Test Set Metrics for Coords + Geologic PCs + Tessera PCs GP

	Metric	RMSE		NSE		KGE	
Model	Horizon	Median	IQR	Median	IQR	Median	IQR
Baseline	H1	1.882	2.480	-74.892	467.421	0.553	1.005
	H12	1.851	2.854	-87.226	419.191	0.050	0.602
N-HiTS	H1	1.860	2.790	-79.911	453.317	0.548	0.787
	H12	1.868	2.831	-80.745	400.545	0.475	0.894
C2 ZS	H1	1.989	2.532	-76.023	412.634	0.542	0.873
	H12	1.810	2.781	-81.224	384.082	-0.001	0.951
C2 FT	H1	1.829	2.543	-88.082	442.865	0.449	0.991
	H12	1.879	2.811	-79.482	404.308	0.485	0.789

	Metric	MPIW		PICP		Winkler	
Model	Horizon	Median	IQR	Median	IQR	Median	IQR
Baseline	H1	18.865	6.045	1.000	0.000	18.901	6.154
	H12	19.248	6.542	1.000	0.000	19.304	6.480
N-HiTS	H1	19.222	6.550	1.000	0.000	19.277	6.472
	H12	19.200	6.532	1.000	0.000	19.259	6.480
C2 ZS	H1	18.524	6.235	1.000	0.000	18.551	6.402
	H12	19.199	6.306	1.000	0.000	19.238	6.328
C2 FT	H1	18.758	6.189	1.000	0.000	18.793	6.178
	H12	19.175	6.565	1.000	0.000	19.233	6.471

D.3. Coupled Spatiotemporal

Additional results for the Coupled Spatiotemporal models are outlined in this section.

D.3.1. Kronecker Gaussian Process Performance Metrics

Tables of metrics for the Kronecker Gaussian Process are contained here.

Table D.12.: Kronecker GP Test Set Performance Metrics — Median [IQR]

Model	RMSE	NSE	KGE	MPIW	Winkler	PICP
KGP No-PC	1.611 [2.806]	-57.062 [381.220]	-0.720 [0.374]	31.355 [60.424]	36.459 [61.880]	1.000 [0.000]
KGP PC	1.594 [2.205]	-69.583 [262.673]	-0.694 [0.297]	10.711 [7.606]	12.123 [10.415]	1.000 [0.000]

D.3.2. Cross Attention Performance Metrics

Performance metrics for the cross attention models are contained here.

Table D.13.: Attention Metrics for Test Set (Horizon 1) — Median [IQR]

Model	RMSE	NSE	KGE	MPIW	Winkler	PICP
Attention	1.440 [2.022]	-48.446 [238.638]	-0.752 [0.400]	1.370 [0.626]	66.556 [164.141]	0.000 [0.513]
GP-Attention	1.540 [2.155]	-51.327 [226.060]	-0.415 [0.005]	1.592 [0.000]	74.494 [187.300]	0.000 [0.893]

Table D.14.: Attention Metrics for Test Set (Horizon 12) — Median [IQR]

Model	RMSE	NSE	KGE	MPIW	Winkler	PICP
Attention	1.519 [1.931]	-52.361 [214.742]	-0.516 [0.308]	1.383 [0.704]	75.383 [158.994]	0.000 [0.597]
GP-Attention	1.539 [2.189]	-48.848 [217.426]	-0.415 [0.004]	1.645 [0.000]	71.911 [189.777]	0.000 [0.923]

D.4. Best Models

This section contains additional analyses related to the Best Spatiotemporal models section.

D.4.1. Maps & PC-Space Plots

The performances on maps and principal component space are contained here.

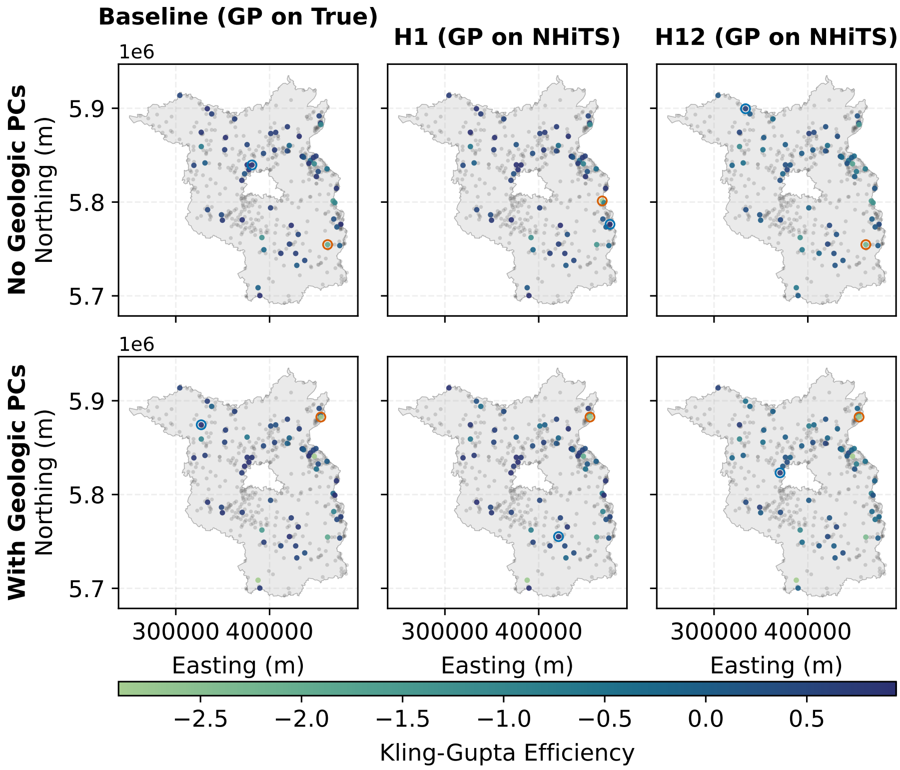

Figure D.1.: Geographic maps with KGE at Test-Test holdout wells and training wells depicted in grey. Poorest performing wells are along the Oder river.

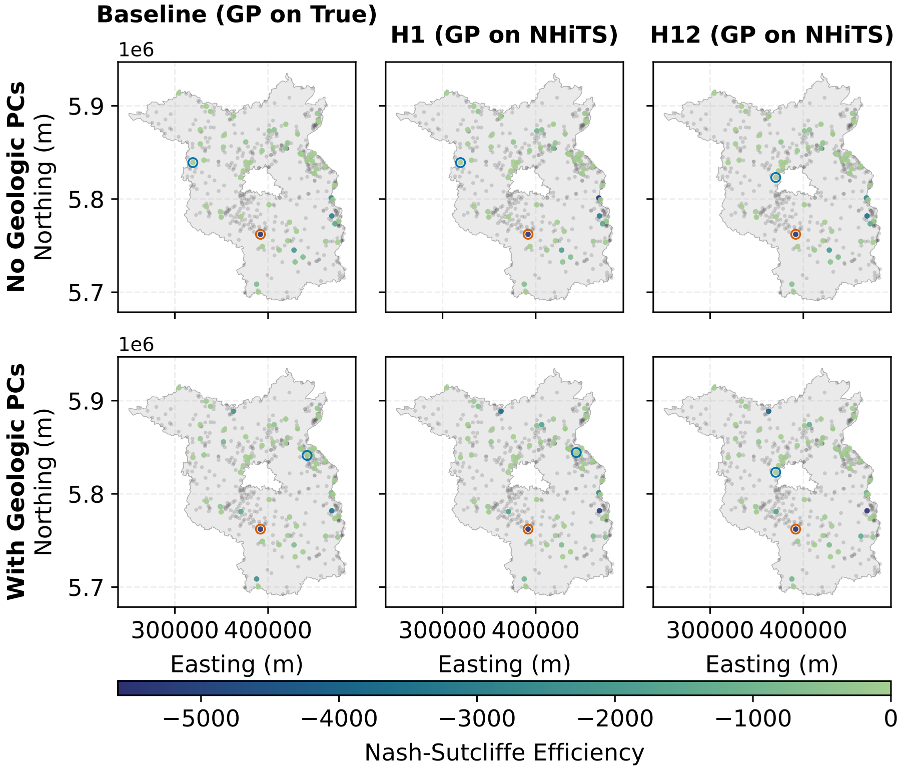

Figure D.2.: Geographic maps with NSE at Test-Test holdout wells and training wells depicted in grey.

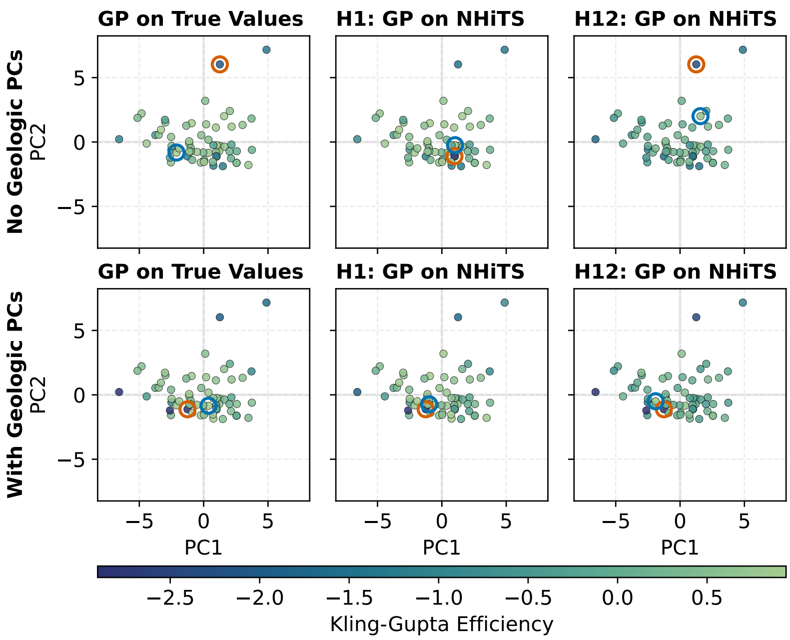

Figure D.3.: Principal component space with KGE at Test-Test holdout wells. GPs with only coordinates appear to have worse performance at wells with very large values of PC2.

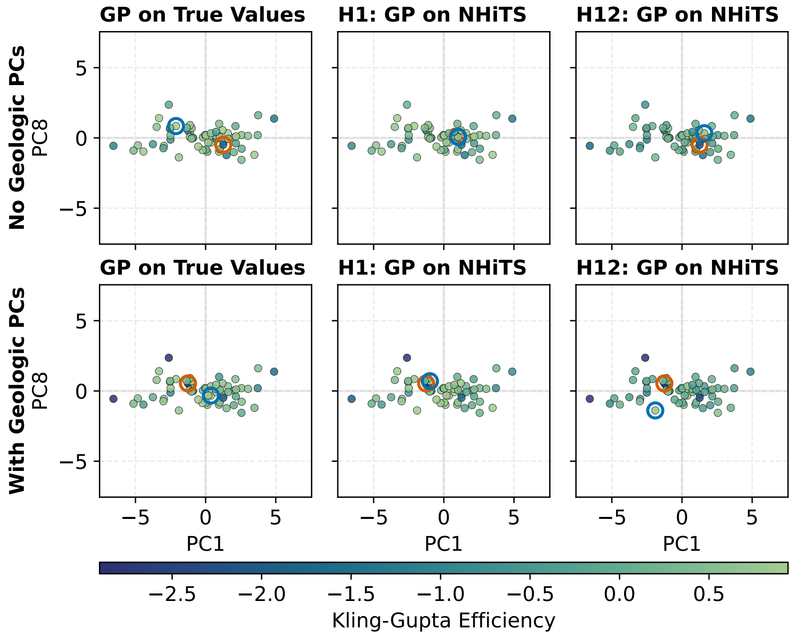

Figure D.4.: Principal component space with KGE at Test-Test holdout wells. No clear pattern except extremes in the space tend to have worse performance than those in the center.

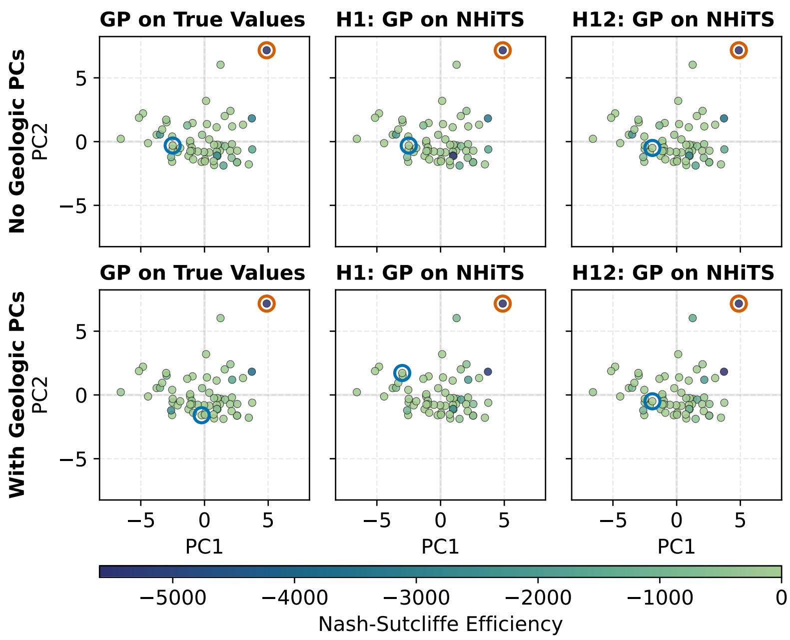

Figure D.5.: Principal component space with NSE at Test-Test holdout wells.

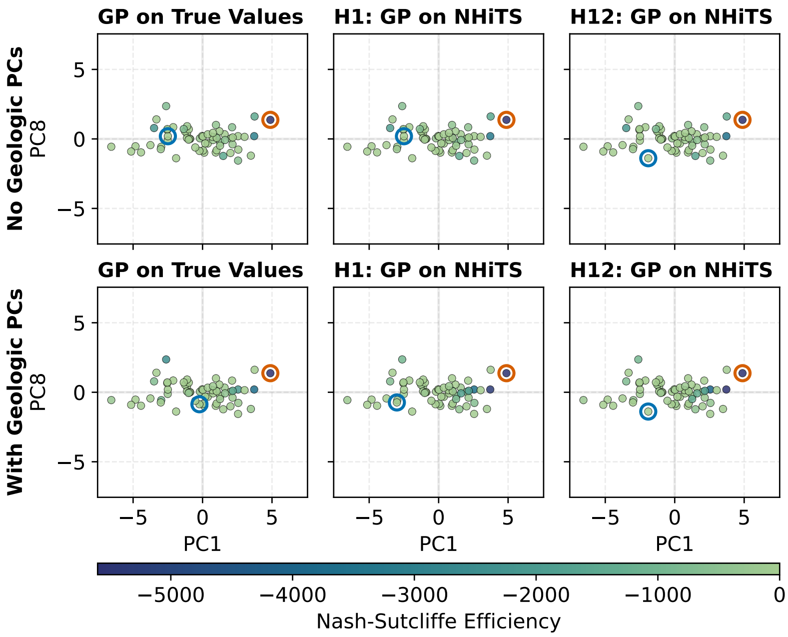

Figure D.6.: Principal component space with NSE at Test-Test holdout wells.

D.4.2. Model Uncertainty

In order to determine when the decoupled Gaussian process on N-HiTS forecasts model's point predictions are trustworthy, it is worth looking at the forecast intervals and the model's uncertainty with the predictions found in the predicted standard deviation. The predicted standard deviation gives an estimate of the spread of the possible functions around the inference point. To this end, we categorize the possible outcomes here:

1. The predicted standard deviation is low and the residual is small (desired behavior)
2. The predicted standard deviation is high and the residual is large (desired behavior)
3. The predicted standard deviation is low and the residual is large (overconfident)
4. The predicted standard deviation is high and the residual is small (underconfident)

Points 1 and 2 are the desired behavior since the model predictions would be trusted if the predicted standard deviation is small and not trusted if the predicted standard deviation is large. Points 3 and 4 denote cases where the predicted standard deviation cannot be used as a proxy for model trust. While this is one possible way to better understand when to trust the model predictions, more work is necessary and other directions could be taken.

To better characterize the model's performance, we explore plots showing the relationship between predicted standard deviation and the residuals and for brevity, we consider the Gaussian process with geologic principal components on N-HiTS forecasts on the Test-Test data. In Figure D.7 we plot the predicted standard deviation against the residuals and absolute residuals. While the residuals appear to have a mean near zero, there is some correlation between the predicted standard deviation and both residuals and absolute residuals. Although it is not a perfect predictor of the residuals, there are sites near the origin with both small predicted standard deviation and small residuals in all plots considered. In Figure D.8 histograms of the correlations between predicted standard deviation and the residuals grouped by site are shown. In the horizon 1 plot, there appears to be a mass of sites near a correlation of 0.5 to 0.75 indicating a linear predictor could be used to estimate the residual at unlabeled sites with similar characteristics to these with moderate performance. In the horizon 12 plot, there appears to be a mass of sites from approximately -0.3 to -0.75 indicating a linear predictor could also be used here. The mass shifted between the horizon 1 and horizon 12 plots from positive to negative which could be due to the behavior of the N-HiTS forecast, going from one side of the true series to the other thus changing the sign of the residuals.

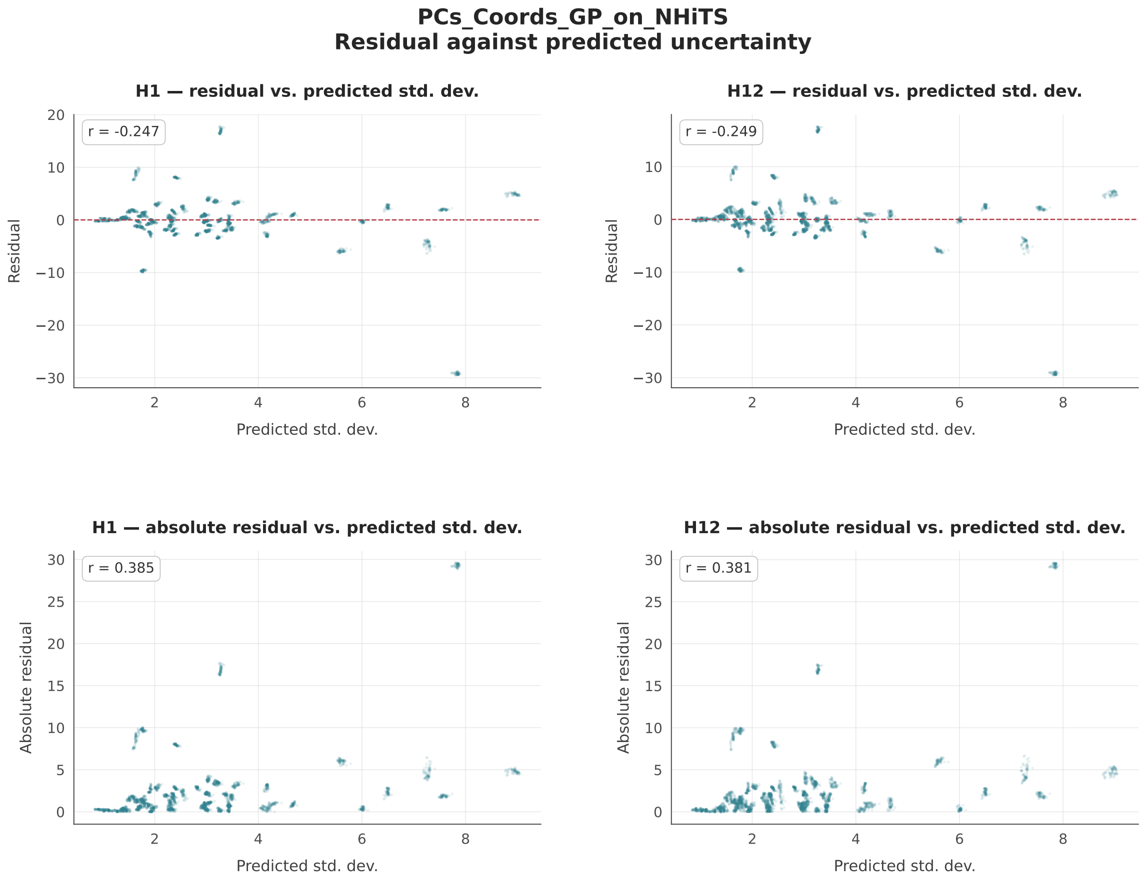

Figure D.7.: Global correlation and scatter plot of residuals against predicted standard deviation. One can see there is some trend though the residuals are primarily centered around 0. The largest residuals tend to have larger predicted standard deviations though not in all cases.

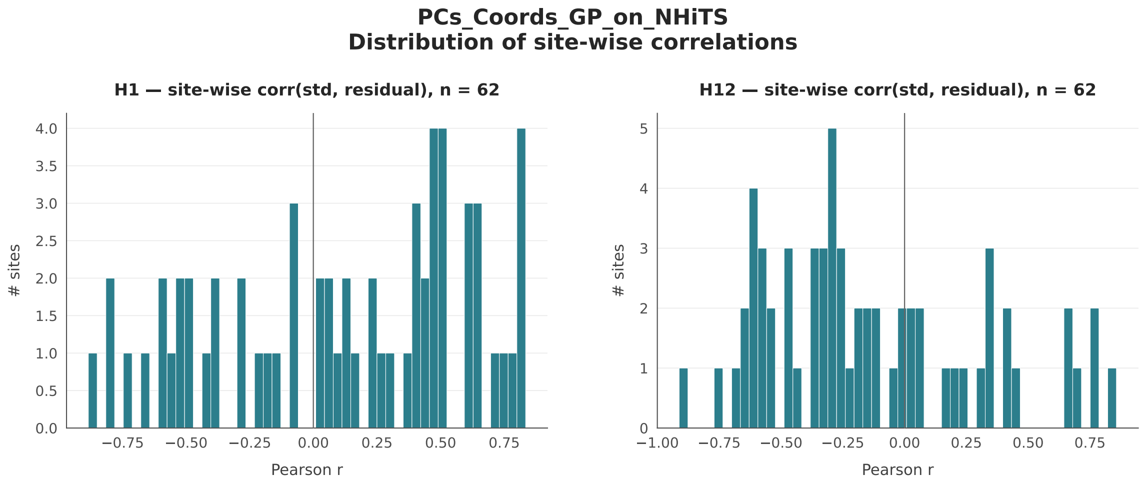

Figure D.8.: Histograms of the site-wise correlations of the predicted standard deviation and residuals. Larger correlations indicate sites where large residuals co-occur with large predicted standard deviations.

D.4.3. Relative Comparisons

More plots on the relative comparisons are outlined in the following

Best Model Against Baselines These plots look at the relative comparisons between the baselines and forecast models.

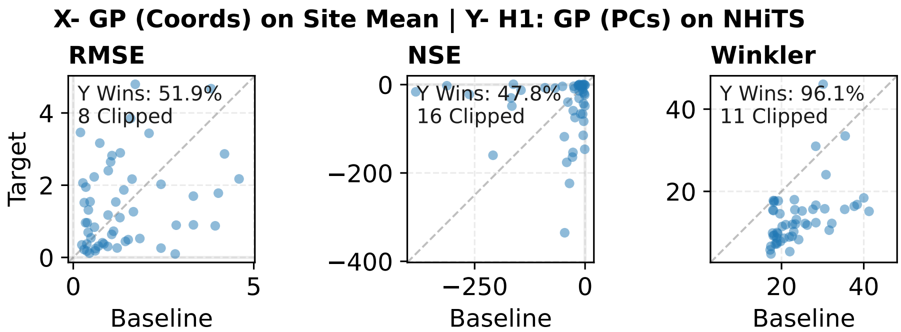

Figure D.9.: Coordinate only GP on historical means vs PCs GP on H1 N-HiTS forecasts. GP on N-HiTS forecasts wins in RMSE and Winkler in the majority of Test-Test wells. This indicates our model is better than a simple coordinate-only, historical data spatial interpolator at a short horizon.

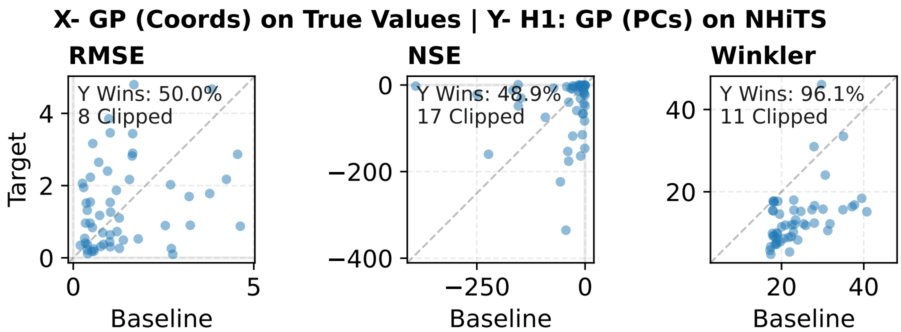

Figure D.10.: Coordinate only GP on true values vs PCs GP on H1 N-HiTS forecasts. GP on N-HiTS outperforms in the majority of cases in Winkler and has the same number of wins in RMSE. This indicates the PCs aid in distributional forecasting and interval predictions.

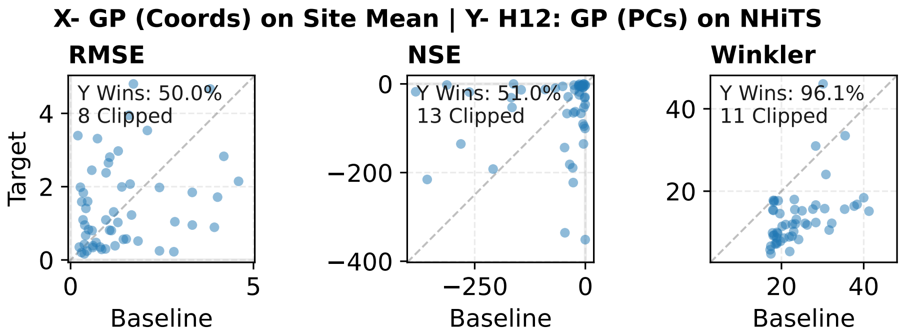

Figure D.11.: Coordinate only GP on historical means vs PCs GP on H12 N-HiTS forecasts. GP on N-HiTS wins in the majority of cases in Winkler and half of cases in RMSE. This indicates our model is better than a simple coordinate-only, historical data spatial interpolator at a horizon of 12 weeks.

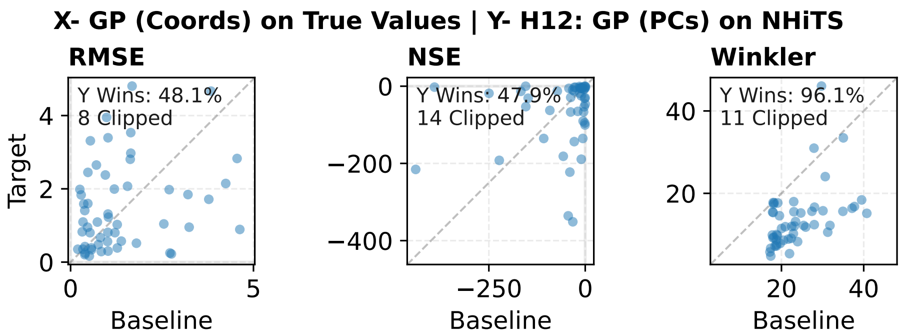

Figure D.12.: Coordinate only GP on true values vs PCs GP on H12 N-HiTS forecasts. GP on N-HiTS wins in the majority of cases in Winkler and nearly ties but loses in RMSE.

Ablation of Geologic PCs The plots here examine the differences between the forecast models with and without principal components

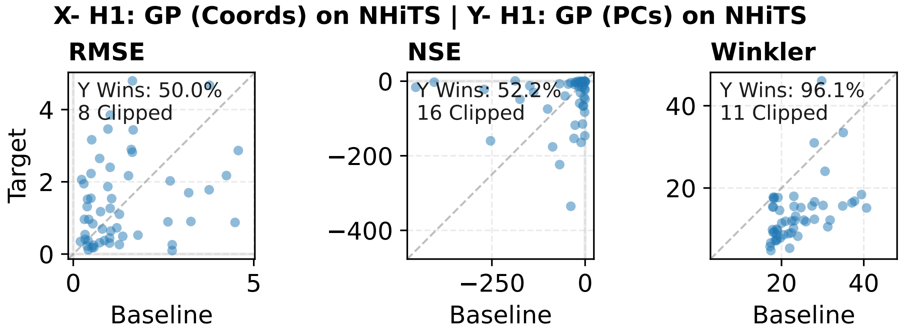

Figure D.13.: Coordinate only GP on H1 N-HiTS forecasts vs PCs GP on H1 N-HiTS forecasts. The methods tie in wins for RMSE, the PCs GP outperforms in Winkler. The win proportions are identical to the coordinate only GP on true values, indicating the spatial model is the limiting factor.

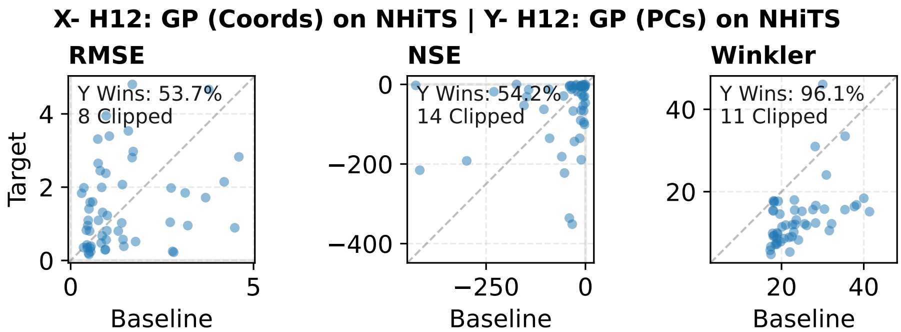

Figure D.14.: Coordinate only GP on H12 N-HiTS forecasts vs PCs GP on H12 N-HiTS forecasts. The PCs GP outperforms in the majority of cases for RMSE and Winkler. The performance is closer than that of the GP fit on true values showing degredation of the forecast quality as the horizon increases and therefore propagation of errors into the GP.