Evaluating Streamflow Prediction in Highland versus Fenland Catchments: A Machine Learning Approach

Aline Van Driessche

Supervised by:
Prof. Emily Shuckburgh
Dr. Robert Rouse

Department of Computer Science

Department of Earth Sciences
University of Cambridge
United Kingdom
28/06/2024

Declaration

This report is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text and/or bibliography.

Code and Data Availability

All code used in this research is hosted on the following public GitHub repository : `https://github.com/AlineGenvision/Lowlands-vs_highlands-streamflow`. An archived version of the repository is also available on Zenodo: `https://zenodo.org/doi/10.5281/zenodo.12581614`. Raw ERA5 (and ERA5-land) outputs can be downloaded from COPERNICUS' climate data store: `https://cds.climate.copernicus.eu/cdsap`; all entries used are detailed in the Appendix (table 6). Additionally, streamflow outputs per catchment are accessible from the NRFA National River Flow Archive: `https://nrfa.ceh.ac.uk/data`.

Acknowledgements

I would like to thank my day-to-day supervisor, Robert Rouse, for giving me the opportunity to learn so many new things in such a short timeframe and for his constant encouragement and insightful feedback. I am also greatful for the general oversight provided by Prof. Emily Shuckburgh, their combined support was vital to the success of this project. Additionally, I am grateful to the AI4ER support staff, Annabelle Scott and Adriana Dote, for creating a supportive environment and providing the necessary encouragement throughout the project.

Abstract

Climate change significantly impacts hydrology, altering precipitation patterns, river flows, and water availability. Accurate streamflow predictions are necessary to mitigate flood risks and manage water resources. Directly measuring streamflow is challenging and resource-intensive, hence the rapid evolution of machine learning (ML) models based on measurable inputs like precipitation and temperature. This study examines the performance differences of an existing Neural Network (NN) and Gaussian Processes (GPs) between highland and lowland regions across the UK.

Snowmelt is identified as a critical factor for highland catchments, which was previously unaccounted for. Other experiments reveal that model performance is fundamentally limited by the ERA5 dataset, which has notable biases and limitations, especially in highland regions. GPs underscore the importance of snowmelt and input variable significance but also face challenges in highland areas, highlighting that a model's effectiveness is tied to data quality.

This research emphasizes the need for better representations of highland catchment dynamics and addressing biases in widely used datasets like ERA5. While more accurate, localized datasets can improve model performance, they are impractical for developing a universal model applicable across diverse conditions. Developing such a universal model is crucial to improve streamflow predictions and manage water resources effectively.

Contents

1 Introduction

Hydrology, the study of water in the environment, is increasingly important as climate change alters precipitation patterns, river flows, and water availability [1, 2]. The Intergovernmental Panel on Climate Change (IPCC) latest report indicates a projected rise in number and intensity of extreme precipitation events in all climate scenarios, leading to significantly increased watervolumes [3]. As streamflow is highly correlated with flood risk, accurate streamflow predictions are crucial to enable better preparation and response to potential flooding events [1, 4, 5]. Additionally, compounding factors such as population growth, urbanization, and increased demand for potable water both influence and are influenced by water catchments worldwide [6]. Understanding streamflow patterns is vital for effective water resource management [7], ensuring sufficient water supply during droughts [8] and to optimize systems dependent on (consistent) water flow such as hydroelectric power generation [9].

Measuring streamflow directly is challenging and resource-intensive [10]. Measurement gauges only cover a limited fraction of the stream network globally [11, 12]. Therefore, over the last 50 years, researchers developed various flow estimation methods evolving from physical-theory based systems [13, 14] to Machine Learning (ML) methods such as Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Random Forests (RF) [15, 16, 17]. Many of these models rely on meteorological input variables such as precipitation, temperature, and humidity [11]. Similar variables are often available in predictive climate models, so that the streamflow prediction models can also be applied across future climate scenarios [4].

This research builds on existing work that uses Artificial Neural Networks (ANNs) [18] for streamflow prediction within a single catchment. The developed model is applied to various basins across the UK, highlighting the poorer performance in highland areas compared to lowland regions (fenland). Multiple experiments in this study aim to enhance the streamflow prediction through both model and data improvements. A critical assessment of input feature selection, processing techniques, and dataset choices identifies key areas for enhancement. Additionally, Gaussian Processes are introduced as a new method, offering a probabilistic streamflow modelling approach with significantly greater interpretability than NNs [19]. The parameter-model dependencies help to understand the dynamics between the meteorological input parameters and streamflow predictions [20].

By testing and validating these potential improvements on a variety of catchments, the aim is to develop robust and accurate streamflow prediction models that can be applied in diverse geographic and climatic conditions globally [12]. Understanding the root causes for performance variations across contexts is crucial to interpret the predicted results. With these insights, the models can provide reliable predictions that can inform water resource management, flood risk mitigation, and energy production strategies under changing climate conditions [6, 7].

2 Methods

2.1 Catchments

The National River Flow Archive (NRFA) serves as the focal point for hydrometic data across the United Kingdom [21]. The NRFA provides catchment boundaries and streamflow information, which is used as groundtruth in this study. A catchment is the entire geographical area drained by a river and it's tributaries [22]. For every documented catchment, daily streamflow values are available, measured at gauging stations around the catchment outlet [21].

Following [18], the study focuses on three catchments in detail: the Bedford Ouse at Roxton; the Severn at Haw Bridge and the Findhorn at Shenachie. For a broader comparative study between the lowland and highland areas, modelling is expanded across 23 other catchments that vary in size and geographical location (see figure 1).

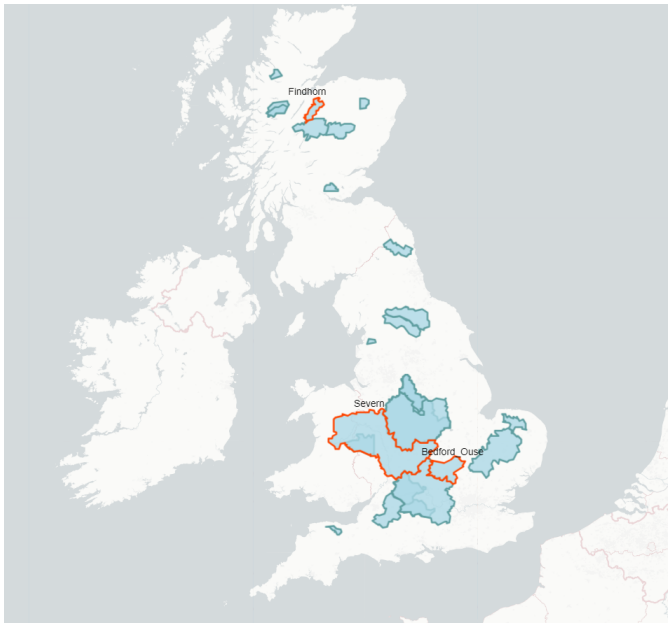


Figure 1: Overview of the extent of the 26 studied catchments across the UK (boundaries as retrieved from the NRFA). The three focus catchments for this study are indicated in red.

2.2 Meteorological Variables

Meteorological input variables are used to account for the water entering and leaving the catchment system. Water enters the system in the form of precipitation and then either evaporates (driven by temperature and humidity), is stored in the catchment system (reflected in soil moisture variables), or contributes to the streamflow output (target variable) [22]. TWind speed is also considered, as it influences the rate of these transition processes. Previous studies, such as [18, 20, 23], confirm the choice of above-defined input variables. All these variables are available as ECMWF (European Centre for Medium-Range Weather Forecasts) Re-Analysis products, fifth generation (short: ERA5 data) [24]. ERA5 is a global climate and weather reanalysis product (combining model data and observations), available at a 0.25°x 0.25° grid resolution. For this research, hourly data points are retrieved from ERA5 over the period 1980-2020 through the Copernicus Climate Change Service (2023) [25], see 5. As the models are catchment-specific, static input variables such as topography or soil type are not considered (they should be internalised by the models)[20].

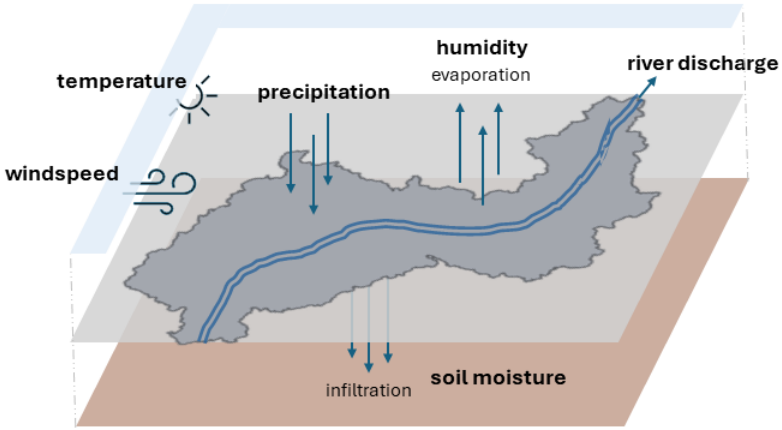
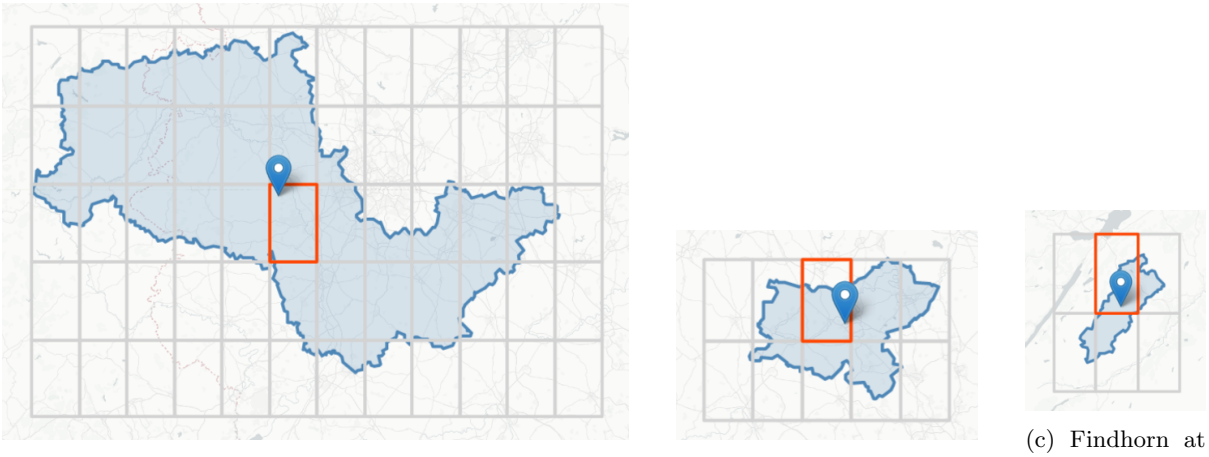


Figure 2: Schematic overview of the hydrological processes within one catchment system. Input variables used for training are in bolt.

2.3 Data Preprocessing

The hourly ERA5 data points are converted to daily input variables for model training. Temperature, humidity, windspeed, and soil moisture are approximated using the value at 12:00 PM. However, for precipitation, the precipitation values are accumulated over 24 hours, as the total volume is more critical than a single hourly measurement [22]. The 24 hours aggregation is performed from '9to9', in accordance with the streamflow measurements from NRFA. The NRFA daily measurements refer to a 'water-day', which stretches from 9am to 9am, instead of a calendar day [21].

Spatially, the input values are resized from the ERA5 0.25° by 0.25° grid to a single value per parameter for the entire catchment. This value is obtained through bilinear interpolation, following equation 9 at the catchment's centroid (see figure 3 and further elaboration in 5.3) [18].



(a) Severn at Haw Bridge

(b) Bedford Ouse at Roxton

(c) Findhorn at Shenachie

Figure 3: Catchment boundaries on the ERA5 0.25° by 0.25° grid. The red boxes indicate the four nearest corners to the centroid, which are used for linear interpolation of the input values.

Lastly, in accordance with [18], the 30, 90, and 180-days rolling means of temperature and precipitation are used as proxies for soil moisture, representing the long-term water storage of a catchment. Reducing the number of input variables or using only commonly available ones is desirable to develop a model that can be applied globally [12].

2.4 Training Details

2.4.1 Neural Networks

The NN research expands on the existing application from [18], where a simple feedforward NN is used to predict the streamflow in a single catchment. The model is an MLP [26] with two hidden layers containing respectively 64 and 16 neurons, using SiLU [27] activations. The most effective optimiser is AdamW [28], minimising the RMSE (equation 1) on a learning rate of 10^{-3} and no regularisation. Further details on the network are included in the Appendix 5.4.

All models use a train/validation/test split: the training data stretches from 1980-2009 and testing (depending on availability) from 2010-2020. Early stopping on the validation set is used to avoid extreme overfitting (with maximum 9000 training iterations). Inputs to the model are normalised, using values from the training set only, to help faster convergence and prevent exploding gradients [26]. All models are trained on consumer hardware utilising GPU acceleration where possible. This is carried out through the PyTorch [29] deep learning framework.

2.4.2 Gaussian Processes

Another method used in this research are Gaussian Processes (GPs). A GP is a probabilistic method that works similar to fitting a distribution to a dataset, but it deals with entire functions (kernels)[30] instead of points. GPs are based on prior beliefs about the function it tries to model (e.g. the relationship between rainfall and streamflow). Using measurement data, the prior belief updates to the posterior

distribution. This distribution allows to make predictions about new, unseen datapoints. Importantly, the prediction also indicates a range of certainty about that prediction [19]. The design of the kernel (see Appendix 5.4) is crucial, as it provides the basis of the distribution. The kernel's parameters, such as lengthscales and variance define how the model interprets the relationships and variability within the data. The lengthscale determines the extent to which points influence each other, whether the function varies smoothly or rapidly, while the variance measures the overall amplitude of variations in the data [31].

GPs quickly become computationally expensive as training involves operations on an $n \times n$ covariance matrix, scaling cubically with the number of data points [32]: $\mathcal{O}(n^3)$. As the streamflow training data spans over 30 years with multiple features per daily input, training time increases rapidly. In order to avoid such computational complexities, Sparse Stochastic Variational Gaussian Processes (SSVGPs) are used to approximate the posterior of the GP. They represent the distribution with a smaller set of points ('inducing points') than the original, large input dataset [33]. These inducing points (m) are initialized randomly and optimized during training with stochastic gradient descent [34]. This reduces the computational complexity to $\mathcal{O}(nm^2)$.

All the GPs are trained with the same train/test split as the ML models, and implemented with GPJax, a Gaussian Process framework in JAX [35]. Using the SSVGPs makes it computationally feasible to train on customer hardware, further details in Appendix 5.

2.5 Evaluation Metrics

Two standard regression metrics are used to evaluate ML model performance: Root Mean Squared Error (RMSE) and the Nash-Sutcliffe Efficiency (NSE). The RMSE ranges between $[0, \infty]$, with better results closer to 0. However, as the flow ranges of different catchments vary heavily in magnitude, comparing their RMSE scores directly would be misleading. Therefore the Nash-Sutcliffe Efficiency (NSE) is used, as comparison metric. NSE is a normalisation of the Mean Squared Error (MSE) metric and results in values within the interval $[-\infty, 1]$: 0 corresponds with a mean prediction for all values, while a score closer to 1 indicates increasingly better predictions [13].

If $y_p(t)$ represents the predicted streamflow value at time step t and $y_m(t)$ is the measured value at timestep t, with n timesteps in total, then the evaluation metrics are defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_p(t) - y_m(t))^2} \tag{1}$$

$$\text{NSE} = 1 - \frac{\sum_{t=1}^{n}(y_p(t) - y_m(t))^2}{\sum_{t=1}^{n}(y_m(t) - \overline{y_m})^2} \tag{2}$$

3 Results

3.1 Replication of ANN results

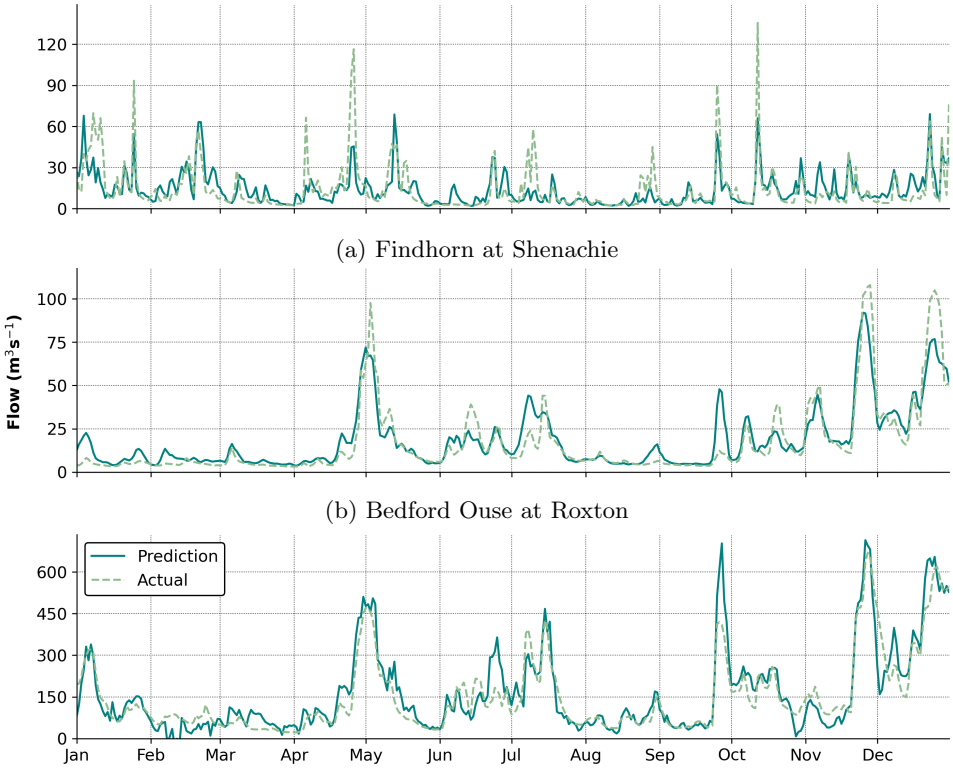
Replication of the methodology in [18] establishes the baseline results to compare the experiments with.

3.1.1 Model Performance

For accurate reproduction, training is performed on data from 1980-2009 and testing on data between 2010 and 2020. Additionally, a 5-fold cross-validation with random train-test splits (retaining 10 years for testing), mitigates concerns about potential bias from the fixed train-test split. Both methods report comparable results to [18], the small reproduction differences can be attributed to the slightly adjusted dataprocessing, (see 2.3).

Site	fixed split RMSE↓	fixed split NSE↑	5-fold validation RMSE↓	5-fold validation NSE↑
Bedford Ouse, Roxton	6.08	0.81	6.87	0.78
Severn, Haw Bridge	46.86	0.84	49.22	0.82
Findhorn, Shenachie	11.05	0.64	11.15	0.64

Table 1: Model performance on the test set (fixed split ranging from 2010-2020) using a MLP with two hidden layers on three individual catchments.



(a) Findhorn at Shenachie

(b) Bedford Ouse at Roxton

(c) Severn at Haw Bridge

Figure 4: Streamflow predictions from the MLP compared to the groundtruth for three catchments in 2012.

Significant performances differences are observed between the three sites, indicating that the learned dependencies are spatially independent. While all models predict average streamflow relatively well (see figure 4), they struggle to accurately estimate peak values. Especially the model trained on the Findhorn catchment consistently underestimates high peak flow events by about half of their actual value.

3.1.2 Expansion across UK Catchments

To elaborate on performance differences, the same MLP is tested on 23 other catchments across the UK. These catchments are selected from the NRFA collection based on data availability (1980-2020) and catchment characteristics: the resultant set represents various sizes, altitudes and geographical locations (an overview is available in Appendix, 5.2).

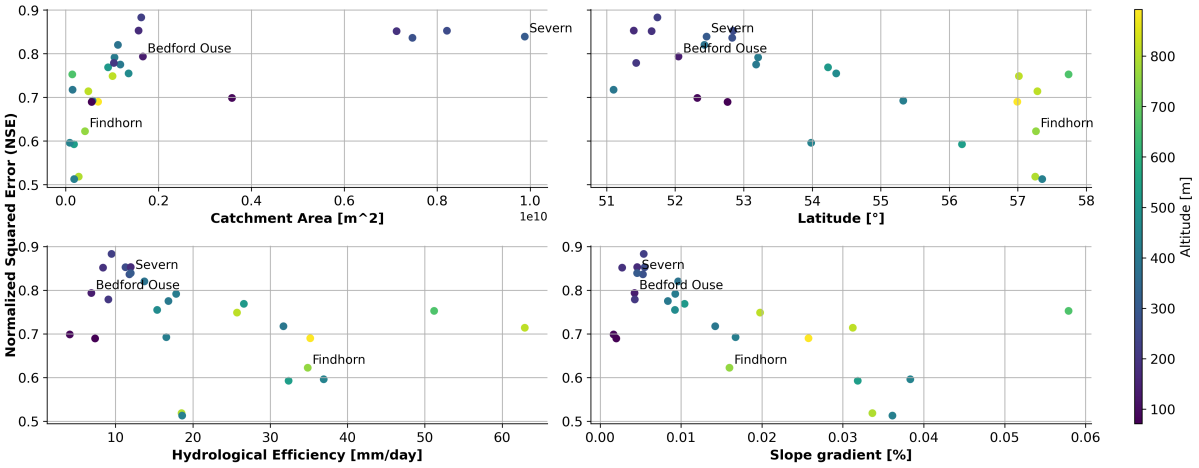


Figure 5: Model performance (NSE) per catchment (26 catchments in total) compared to their catchment characteristics (area, slope gradient, hydrological efficiency and geographical location). The colors refer to the altitude of the catchment.

Figure 5 illustrates a significant variation in model performance across different catchments. The graphs do not conclusively identify the determining factors for these differences, but some patterns are evident. For small to mid-size catchments, there appears to be a correlation between model performance and catchment size. This becomes more apparent when considering Hydrological Efficiency (flow divided by catchment area) [22]. Smaller, 'flashier' catchments (responding quickly to precipitation events) are more challenging to model compared to larger catchments with slower response times. The slope gradient supports this observation: catchments with steeper slopes tend to have poorer model performance. The response time of a catchment also depends on other factors such as soil composition or altitude, both linked to the catchment's geographical location [36]. The plot comparing NSE performance with catchment's latitude reveals a clear linear trend and moreover, it also illustrates the strong correlation between latitude and altitude within the UK. Higher latitudes correspond to higher altitudes (Scottish mountains). Catchments at higher latitudes, particularly in Scotland, perform significantly worse, suggesting that the models are sensitive to specific regional characteristics.

Building on the above-formulated hypotheses, this research presents five novel experiments that aim to improve model performance across UK catchments. These experiments involve using higher temporal resolution data, snowmelt, additional data sources beyond ERA5, another interpolation method, and a higher spatial resolution data.

3.2 Experiment 1: Higher temporal resolution data

The first experiment investigates if the deteriorated model performance in the highlands results from failure to capture the rapid hydrological dynamics of these areas. As shown in figures 1 and 5, the catchments in the highlands are predominantly small catchments [37] with a high Hydrological Efficiency (HE), which consistently results in poorer performance. Sensitivity analysis of the model towards the individual daily precipitation values helps to understand this discrepancy better: small perturbations to input variables allow to analyse and scale the corresponding changes in the model output [38], see figure 6. It indicates that smaller catchments, with a higher HE depend more on recent days of rainfall compared to larger, slower systems.

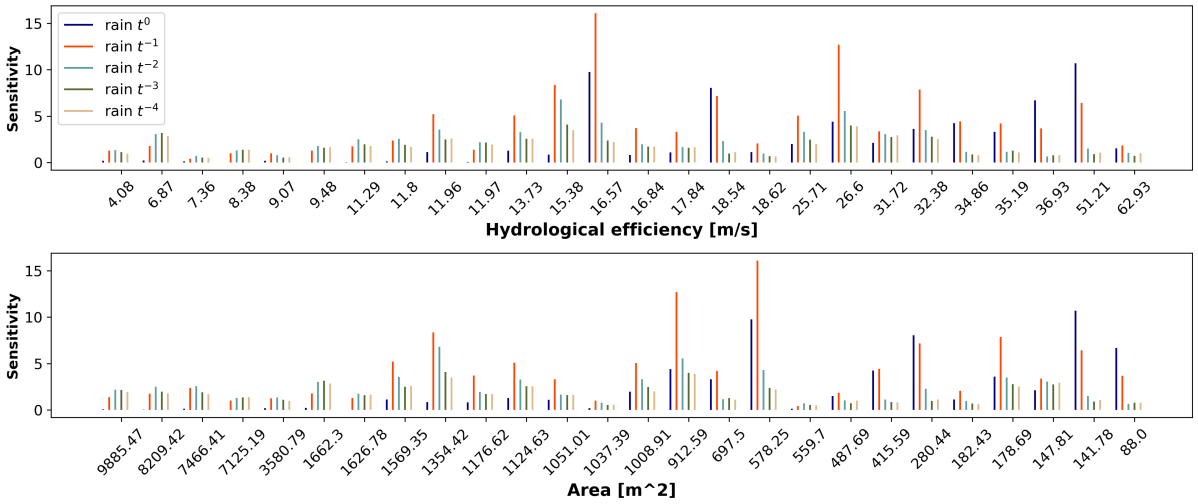


Figure 6: Sensitivity analysis on the precipitation values from the five preceding days in relation to the streamflow prediction day. The sensitivity to these parameters is compared with the catchment's HE and total area across 26 catchments.

Potentially, providing input values with a higher temporal resolution than the so-far used daily scale, gives the model more insights into the exact temporal dynamics from those days. Adding hourly datapoints doubles the number of input parameters, therefore various representations for these points are employed: 2-hourly aggregation, rolling means, and a parameterization method that indicates both the extent and time-scaled occurance of the daily peak value, as visualized in figure 7. Additionally, the model architecture is adjusted to mitigate the increasing amount of input parameters. Tested alterations include increased model capacity (number of layers and size of the layers), hyperparameter tuning (e.g. dropout rate) and L2 regularisation (as weight decay) [26].

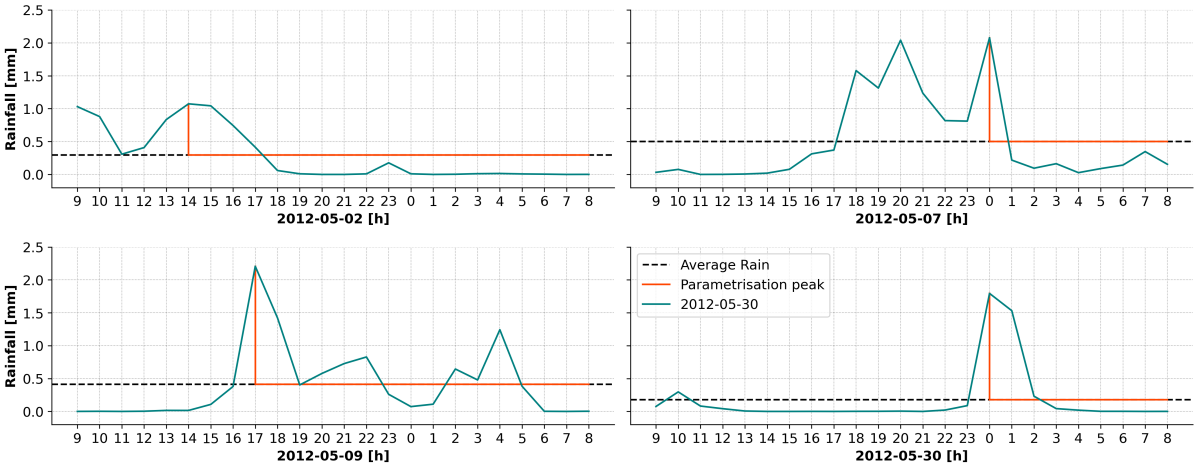


Figure 7: Hourly precipitation values over 24 hours ('9to9'). The red lines indicate parametrisation of the peak value. All the graphs represent dates in May 2012 in the Severn catchment.

NSE ↑	Daily	24 hours	2-hourly	Rolling Means	Parametrisation
Bedford Ouse, Roxton	0.81	0.80	0.79	0.78	0.79
Severn, Haw Bridge	0.84	0.83	0.84	0.84	0.84
Findhorn, Shenachie	0.64	0.64	0.64	0.62	0.63

Table 2: Model performance for when including different variations of hourly data as extra input parameters. Compared with the training only using daily inputs.

Suprisingly, including hourly information does not improve model performance, with neither of the value representations (see table 2) and sometimes performance even goes down (likely due to the additional input parameter complexities) [20]. Since the added hourly values do not provide new insights beyond what is already captured in the daily data, it is safe to conclude that catchment area or HE are not the primary factor affecting model performance in the highlands. However, the heightened sensitivity to recent rainfall events as seen in figure 6 still underscores the need for more accurate and temporally aligned precipitation data to improve model performance.

3.3 Experiment 2: Snow melt data

Another notable difference between higland and lowland areas is the volume of snowfall [39]. Snowfall is considerably more present in higher altitude regions and, due to the colder climate, tends to remain longer before melting. Snow is already included in the current model setup, as the 'total precipitation' feature from ERA5 includes both rain and snow [25], but it assumes that both contribute immediately to the catchment area. However, under freezing circumstances, the snow contribution, in terms of snow melt, experiences a delayed effect [39]. To account for this, the 'snow melt' input parameter is retrieved from ERA5, preprocessed in a similar manner as the precipitation and added to the models inputs. Figure 8 illustrates the correspondance between snow melt and precipitation, highlighting the correlation of the snow peak values with the streamflow.

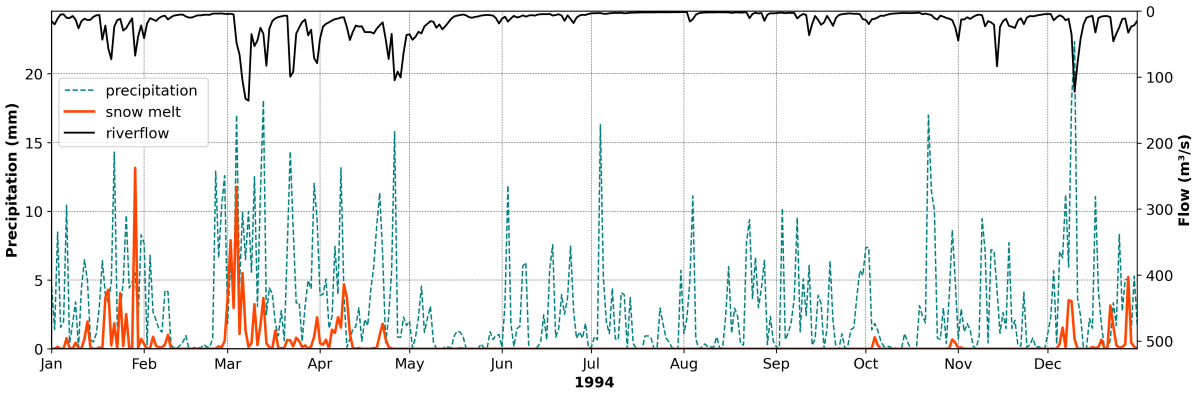


Figure 8: Daily precipitation and snow-melt values (on the bottom) compared to the riverflow values (on the top) in 1994 across the Findhorn catchment.

NSE ↑	Reference	snow melt (P)	snow melt (SM)
Bedford Ouse, Roxton	0.81	0.80	0.80
Severn, Haw Bridge	0.84	0.85	0.84
Findhorn, Shenachie	0.64	0.68	0.71

Table 3: Model performance (NSE) when including snow melt for the Bedford, Severn and Shenachie catchments using either soil moisture variables (SM) or soil moisture proxies (P).

Table 3 makes clear that, as expected, including snow melt improves the model performance a bit in the highland catchment (Findhorn) and maintains the baseline result for others. Furthermore, replacing the soil moisture proxies with actual soil moisture values reveals that precipitation and temperature as proxy alone might be insufficient to accurately represent soil moisture status. Snow also contributes, it provides a gradual and sustained source of water that is critical for recharging ground water [40]. The timing and rate of snowmelt can therefore significantly influence soil moisture dynamics and hydrological processes in a catchment [39]. This experiment certainly proves that snowfall plays a noteworthy role in riverflow prediction and should thus be taken into account, yet it doesn't cover the full gap between high- and lowland prediction performances.

3.4 Experiment 3: Measured precipitation data from NRFA

The next experiment investigates whether the issue lies in a temporal bias within the data rather than the temporal resolution itself. So far all the experiments conducted are based on ERA5, which is a reanalysis product at a relatively coarse resolution (0.25° x 0.25°, corresponding to 20-30kms in the UK) [24]. The ERA5 reanalysis products rely on a combination of observational data and numerical models, which can introduce uncertainties at local scales [41]. This experiment tests model performance with a datasource relying solely on observations: gauged daily precipitation values from the National River Flow Archive (NRFA)[21]. The NRFA provides a catchment-averaged daily rainfall product derived from CEH-GEAR data (1km gridded dataset resulting from observations from the Met Office) [42]. Corresponding to the NRFA river flow values, the daily rainfall is aggregated from '9to9'. As this data is captured from in-situ gauging stations, it ideally reflects the actual hydrological conditions.

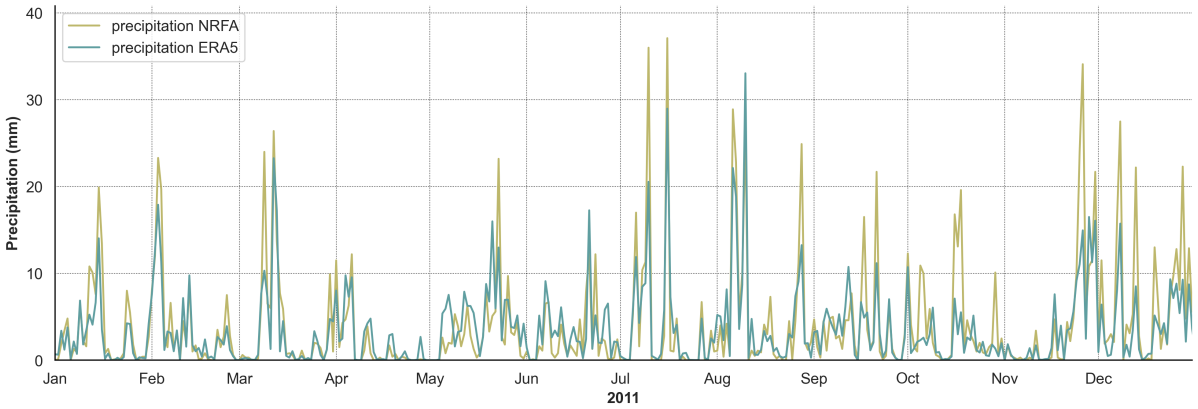


Figure 9: Precipitation values as recorded by ERA5 (reanalysis) versus NRFA (measurements) in 2011 in the Findhorn catchment, see Appendix 5.5 for similar plot in the Bedford Ouse and Severn catchment.

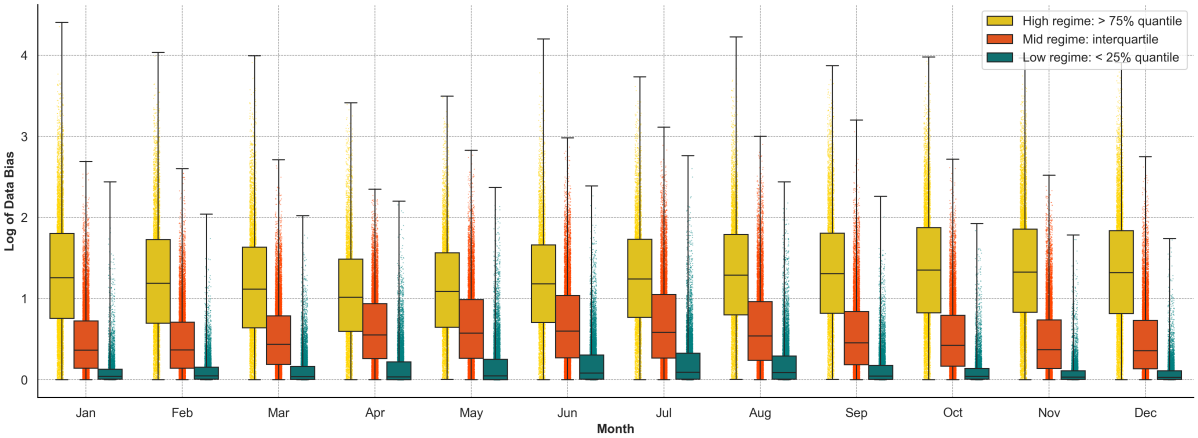


Figure 10: Bias (Normalized Error) in the daily precipitation data (NRFA compared to ERA5) across 26 catchments, biases are log-transformed for better visualisation. The box plots illustrate the data distribution per month, divided into three regimes based on quartiles (high, mid and low).

Figure 9 illustrates that the NRFA data emphasizes precipitation peak values more strongly than ERA5 data. This is further validated by the error analysis in figure 10 which spans 40 years and 26 catchments, and is divided in three regimes based on the error quartiles (percentiles). Although no distinct seasonal or other patterns are evident, the error is significantly higher for the upper regime, indicating that most of the substantial differences lie in the peak values. Since previous experiments struggle to predict peak values, the inherent data bias might be the underlying issue hindering improvements in model performance, as a model's performance is inherently limited by the quality of its data.

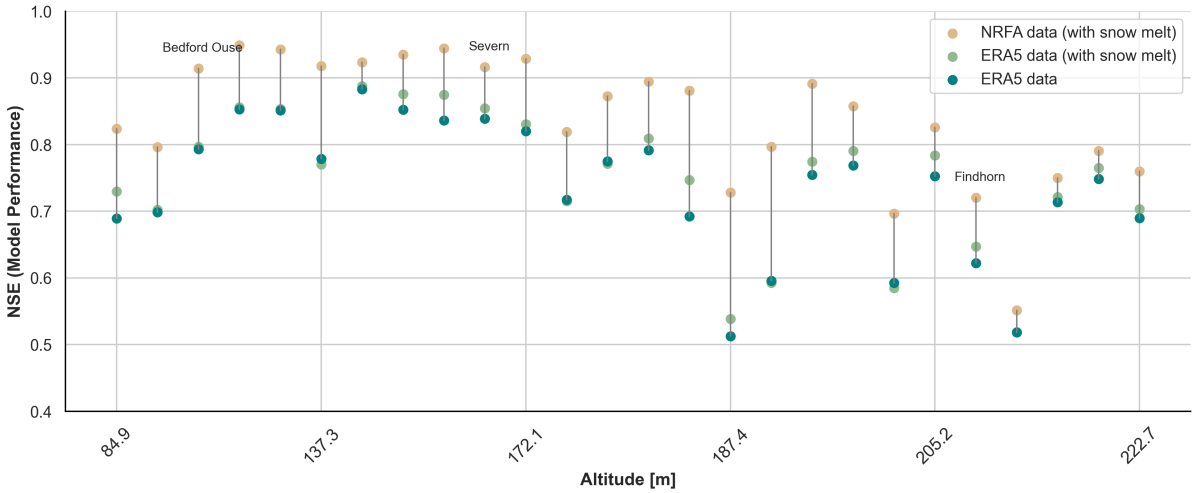


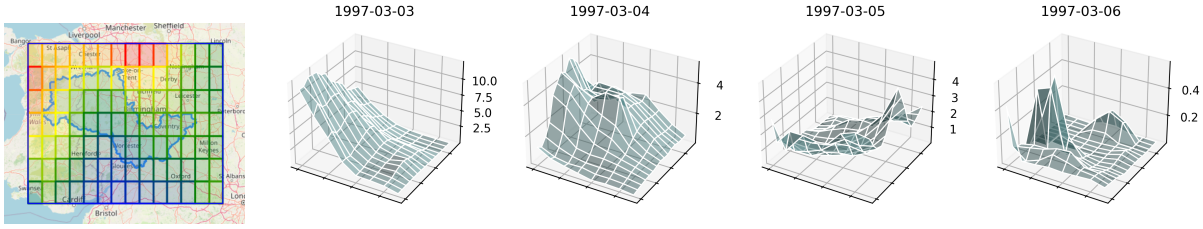
Figure 11: NSE Performance compared across all 26 catchments using Era5 versus NRFA precipitation data, with and without snow melt included (ordered from low to high altitude).

The new model uses the NRFA precipitation values for both the daily rain inputs as the 28-90-180 rolling mean used as a soil moisture proxy (see 2.3), all the other input parameters remain the ERA5 data products. Using the NRFA precipitation inputs improves the performance across all 26 catchments, as shown in figure 11, extending up to 20% improvement for some of the catchments. This significant enhancement proves that accurate precipitation values play a crucial role in streamflow prediction. ERA5 captures the general (low regime) precipitation trends but forms a less reliable foundation for peak values, which are essential for understanding catchment dynamics [43].

However, for developing a broadly applicable model, relying on very localised high-resolution products like the NRFA, while useful for validation, is not ideal [12]. Additionally, even with NRFA data, catchments in the highlands still perform worse than those in lower regions, indicating that the inherent complex dynamics of these catchments are still not fully addressed by using different precipitation data.

3.5 Experiment 4: Surface interpolation method

Expanding on the previous section, it is important to verfiy if the ERA5 data is inherently less accurate than the NRFA data or if the data preprocessing doesn't fully optimize the information contained within ERA5 data. To improve robustness of preprocessing ERA5 input data, this experiment interpolates the data while considering the entire catchment surface, instead of relying solely on the four precipitation values around the centroid (as described in 2.3).



(a) ERA5 grid.

(b) Interpolated precipitation surfaces over 4 days.

Figure 12: The ERA5 0.25°x 0.25° grid resolution for the Bedford Ouse catchment, with the interpolated precipitation surfaces over this grid for 4 days in March 1997. Intersection of the gridlines indicate the available datapoints from ERA5.

As figure 12 shows, using surface interpolated functions increases the presence of finer-grained variability. To calculate the total catchment-precipitation value, the interpolated surface is integrated over the catchment boundary at each timestep (corresponding to one day).To minimize computation time, the integration is approximated by summing the interpolated values at discrete points within the surface catchment S [44]:

$$\iint_S f_p(x,y)\,dA \approx \sum_{k=1}^{N} f_p(x_k,y_k),dA \approx \sum_{k=1}^{N}\left(\sum_{i=0}^{3}\sum_{j=0}^{3} a_{ij}(x_k-x_0)^i(y_k-y_0)^j\right) \tag{3}$$

where:

- $f_p(x,y)$ is the bicubic interpolation formula [45].
- S is the surface over which the integration is performed.
- N is the total number of discrete points (x_k, y_k) within the surface S.
- a_{ij} are the coefficients determined by the function values and the derivatives at the grid points.
- (x_k, y_k) are the coordinates of the discrete points within the surface S.
- (x_0, y_0) are the reference grid points from which the interpolation is performed.

Empirical experiments indicate that a grid resolution of 0.05° is sufficient for this purpose. It captures the necessary detail without the excessive computational burden associated with a finer resolution. This approximate integration ensures that the calculation is both accurate and computationally feasible.

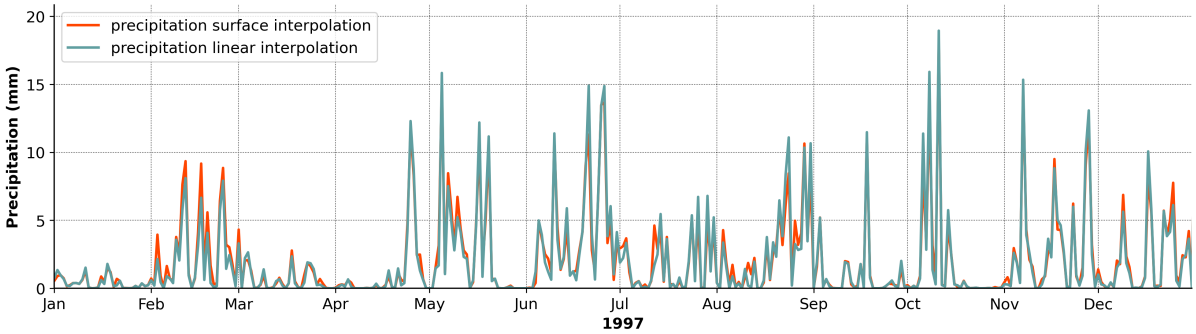


Figure 13: Precipitation values resulting from surface integration over a 0.05° grid for the Severn catchment, compared to the results achieved with linear interpolation around the centroid.

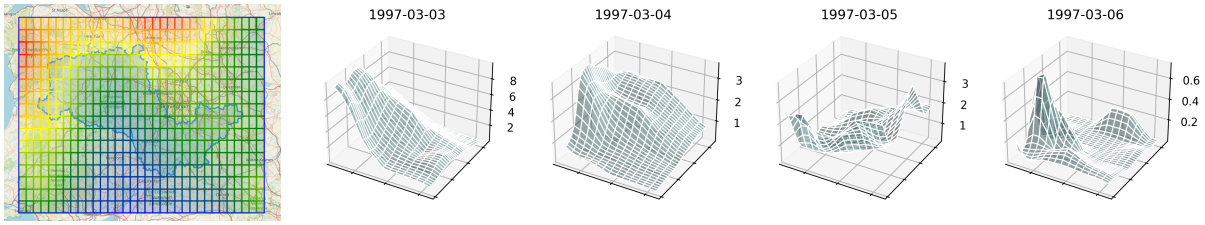
NSE↑	Reference	NRFA	surface interpolation
Bedford Ouse, Roxton	0.81	0.91	0.82
Severn, Haw Bridge	0.85	0.92	0.87
Findhorn, Shenachie	0.65	0.72	0.65

Table 4: Model performance (NSE) using precipitation values obtained through surface interpolation on ERA5 data versus NRFA precipitation values (reference is through centroid interpolation on ERA5 data). All models include snow melt.

According to figure 13, the surface interpolated values capture some of the peaks slightly sharper than the simpler centroid method. However, the variance between surface-based interpolation and centroid-based interpolation results is not as substantial as anticipated. While the surface interpolation method is undoubtedly more robust and slightly more accurate in representing the full catchment [44], the performance improvement is only marginal (see table 4). This suggests that ERA5 data inherently differs from the NRFA data. Even with more robust interpolation techniques, the additional insights gained are limited. It can be concluded that the discrepancies between ERA5 and NRFA data are due to the fundamental differences in the datasets themselves, rather than the preprocessing methods applied. The error is even more pronounced in highland areas, as previous studies have proven that the ERA5 reanalysis is more prone to error in mountaneous areas [43, 46].

3.6 Experiment 5: Higher spatial resolution data

One fundamental difference between the ERA5 data and the NRFA dataset is the resolution: ERA5 data points are available on a grid of 0.25 x 0.25°, (20-30 kms) [20]. In contrast, the NRFA dataset is derived from a 1km x 1 km resolution, providing much more detailed and localized information [21], particularly valuable to capture finer-scale geographical features and local weather [28] phenomena [47]. This experiment uses a dataset with a higher resolution yet globally available: ERA5-land data. ERA5-land offers a higher resolution of 0.1° x 0.1° (approximately 9 kilometers) which potentially allows for improved representation of local processes. When working with ERA5-land data, the daily aggregation (see 2.3) requires careful attention as the accumulation conventions differ from those in ERA5 [48]. The subsequent data processing steps happen in the same way as explained in 3.5: rasterising the data (0.1° resolution), performing the surface interpolation and approximating the integration over the catchment area.



(a) ERA5-land grid.

(b) Interpolated precipitation surfaces over 4 days.

Figure 14: The ERA5-land 0.1°x 0.1° grid resolution for the Findhorn catchment, with the interpolated precipitation surfaces over this grid for 4 days in March 1997. Intersection of the gridlines indicate the available datapoints from ERA5-land.

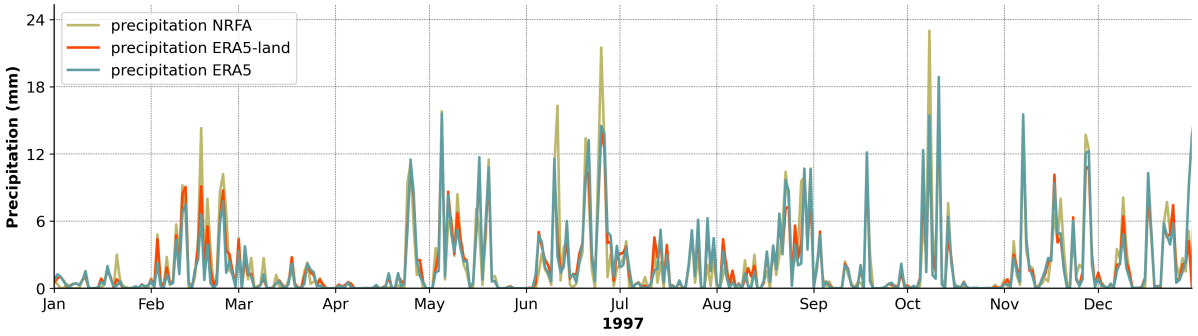


Figure 15: Daily precipitation values obtained from ERA5 and ERA5-land (both processed with surface-based interpolation), compared to NRFA precipitation values in the Severn catchtment in 1997.

The surface-interpolated values from ERA5-land do not deviate much from the lower resolution ERA5 precipitation results (see figure 15), only some peak values are defined a bit sharper with ERA5-land, but they do not follow the bigger differences from the NRFA data. This also reflects in the performance comparison 16. For some catchments (such as the Findhorn), the performance improves a bit, but for the majority using the high resolution data doesn't make any (positive) difference. Across every catchment, the model using NRFA data remains superior. This discrepancy underscores the importance of not only spatial resolution but also the representativity of the underlying data to the hydrological context [49].

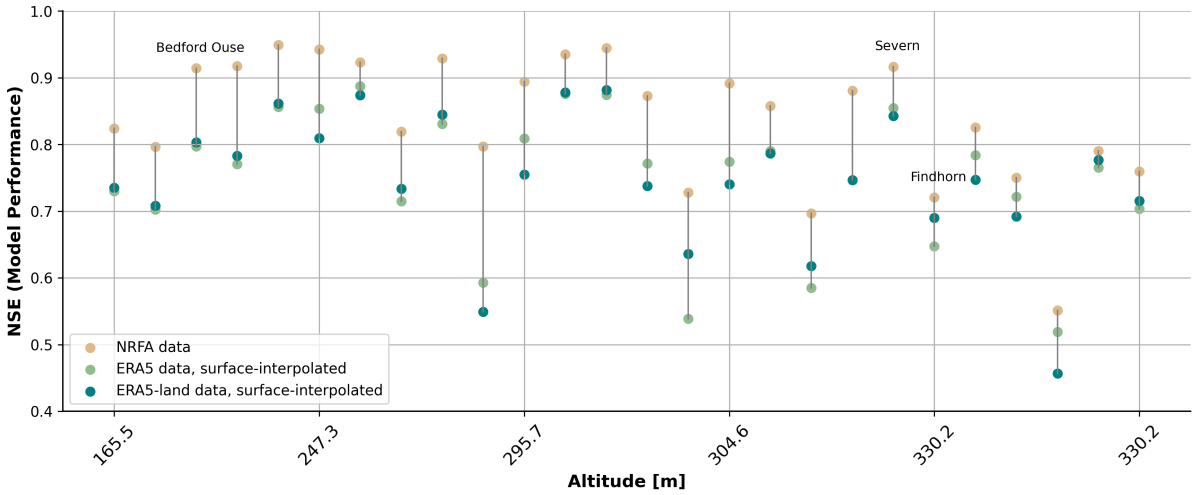


Figure 16: NSE performance compared across all 26 catchment, every vertical line represents an individual catchment, ordered according to altitudes (x-axis). Models based on surface-interpolated precipitation from ERA5 and ERA5-land are compared with the model based on NRFA precipitation.

3.7 Gaussian Processes

The previous experiments prove certain inherent limitations on the data used to train the ANN model. However, also the model itself is worth questioning. This last section zooms in on another method: Gaussian Processes. A Gaussian Process (GP) is a probabilistic machine learning method used for similar tasks as NNs such as supervised regression [32]. Opposed to black-box NNs, GPs provide a way to encode domain knowledge (e.g. incorporate information about the type of correlation between input parameters and the output), through the kernel design (prior) [30]. Design of the kernel for the GP requires careful understanding of the data.

A kernel component that is directly relevant is the Matérn kernel 3/2, as it provides the flexibility to model different degrees of smoothness in the data. This is crucial when dealing with environmental data such as rainfall and streamflow, which exhibit different levels of variability and smoothness over time and space [50]. Additionally, the Matérn 3/2 kernel's flexibility can handle the variability and sudden spikes in the data [51]. Such varying degree of smoothness are defined by the variance parameter σ^2 : when σ^2 =1/2, it becomes the exponential kernel (less smooth) while, when approaching infinity, it becomes the Radial Basis Function (RBF) (very smooth) [52].

$$k_{\text{Matérn-3/2}}(x, x') = \sigma^2 \left(1 + \frac{\sqrt{3}\|x - x'\|}{l}\right) \exp\left(-\frac{\sqrt{3}\|x - x'\|}{l}\right) \tag{4}$$

Given that rainfall is almost linearly correlated with streamflow, the kernel must also include a linear component [19].

$$k_{\text{Linear}}(x, x') = \sigma^2(x \cdot x' + c) \tag{5}$$

To improve the prediction of peak values, an exponential component is added to the prior. This component specifically targets the sharp increases in streamflow during extreme events [53].

$$k_{\text{Exponential}}(x, x') = \sigma^2 \exp\left(-\frac{\|x - x'\|}{l}\right) \tag{6}$$

Lastly, a noise component is included to account for random fluctuations and measurement errors, ensuring robust and accurate predictions [54].

$$k_{\text{White Noise}}(x, x') = \sigma^2 \delta(x - x') \tag{7}$$

Equation 8 shows the final kernel design, the individual components are further documenten in the appendex 5.4.

$$\text{Prior} = (k_{\text{Matern 3/2}} \cdot k_{\text{linear}} \cdot k_{\text{exponential}}) + k_{\text{white noise}} \tag{8}$$

In the interest of reducing computational complexity and effective data use, Sparse Stochastic Variational Gaussian Processes (SSVGPs), see 2.4.2, are used retain the richness of the data while reducing computational requirements [33]. In this experiment, 150 inducing points are randomly distributed over the range of each input dimension, and then optimized to the most informative points during training [34]. Upon training, GPs inherently incorporate the suggested correlations, producing a posterior distribution that reflects the streamflow based on the input variables [19] and predictions can be drawn from this distribution.

Unlike the weights of a NN, which have no direct physical correspondance, the parameters defining GP's kernel components (lengthscale and variance) directly capture information about the data [55]. The variance σ^2 of each kernel component scales its overal contribution to the covariance function [33], summarised in table 5. The high variance of the linear kernel indicates a dominant linear component in the data while also the Matérn 3/2 kernel captures a sigificant portion of the data's variability. The exponential component, while present, does not represent a dominant pattern. Lastly, the low variance of the white noise component suggest that there is a low noise to signal ratio in the input data.

	Matérn 3/2	Exponential	Linear	White Noise	NSE (↑
Bedford Ouse, Roxton	0.135	0.394	0.018	0.001	0.79
Findhorn, Shenachie	0.169	0.398	0.02	0.001	0.60
Severn, Haw Bridge	0.135	0.394	0.018	0.001	0.79

Table 5: Variance parameter for each individual kernel component after training. Compared with the final model performance across three catchments: Bedford Ouse, Findhorn and Severn.

The other important parameter is the lengthscale l (present in both the Matérn 3/2 and exponential component), reflecting how quickly the correlation between points decay with distance (the smoothness of the function) [56]. While each kernel component has a single variance value, the lengthscales are individually adjusted for each input dimension, through Automatic Relevance Detection (ARD) [57]. Smaller lengthscales, as shown in figure 17 suggest that the function varies rapidly with small changes in the rainfall inputs and soil moisture proxies. Larger lengthscales, in the snow melt and some of the temperature parameters, indicate a low likelihood of model changes when the input value changes [58].

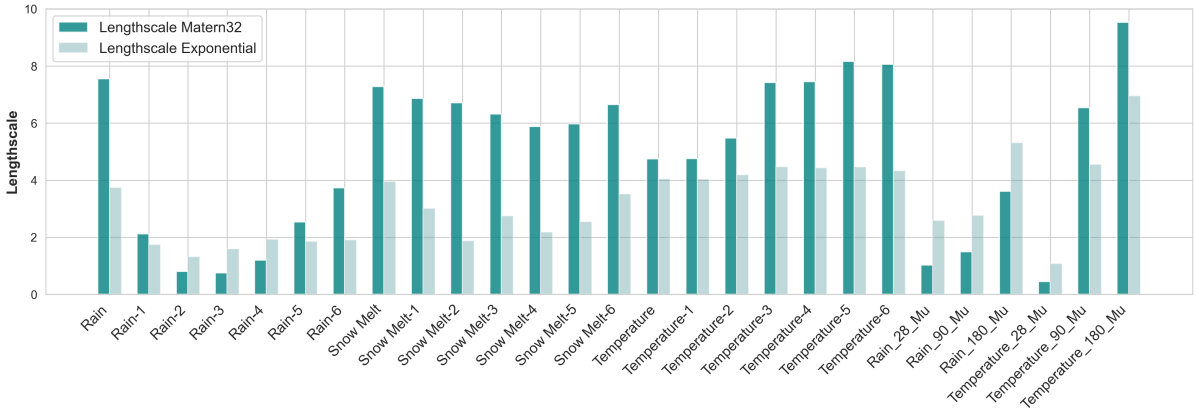
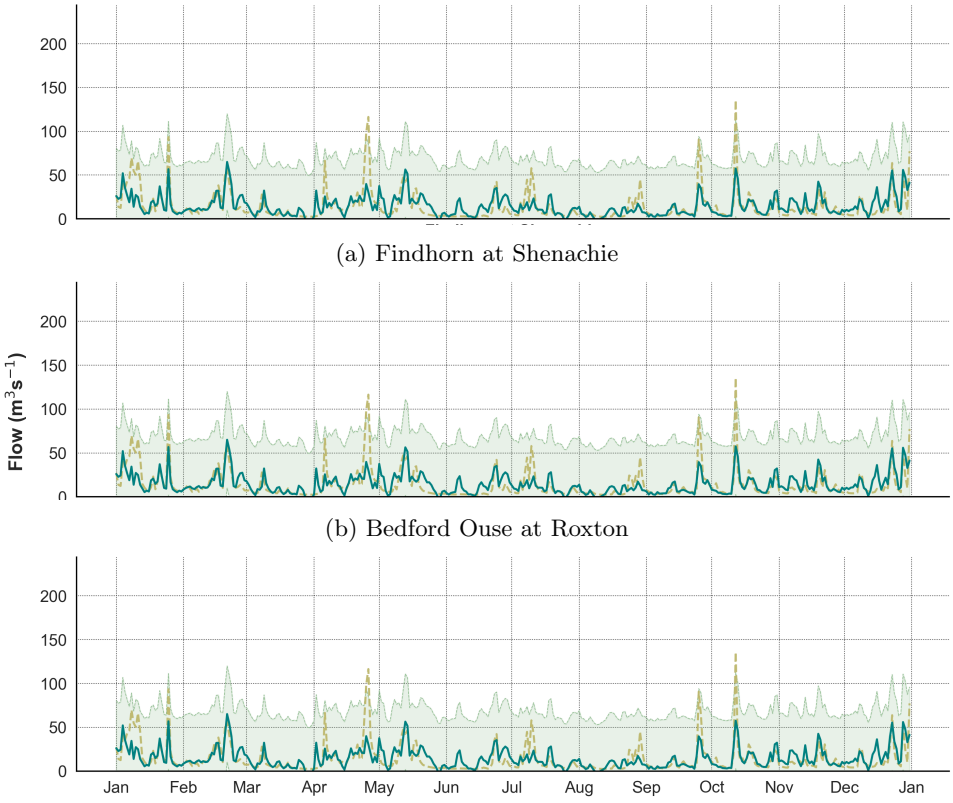


Figure 17: Lengthscale parameter for the Matérn 3/2 and exponential kernel component, on each individual rain, snow, temperature and soil moisture proxy parameter. On the Bedford Ouse, Roxton.

The kernel design and parameter analysis of Gaussian Processes (GPs) already highlight their significantly greater interpretability compared to deep learning methods [55]. Additionally, GPs provide not only predictions but also an estimate of the uncertainty of those predictions [30]. This is crucial for hydrological modelling, where uncertainty can be significant [59]. Figure 18 shows that, for each of the three

catchments, the trained GP effectively predicts the overall trend but struggles to accurately predict the peak values. For the lowland catchments (Bedford Ouse and Severn), the peak values are still within the confidence interval (2 standard deviations) but for the highland catchment Findhorn the actual values are not in reach of the prediction. However, earlier experiments proved that the data sometimes failed to represent values in exactly those areas, which contributes to the underestimation. Lastly, figure 19 concludes that the prediction discrepancies between catchments in high- and lowlands is similar when training GPs as with NNs.



(a) Findhorn at Shenachie

(b) Bedford Ouse at Roxton

(c) Severn at Haw Bridge

Figure 18: Streamflow prediction from trained GP (Matérn 3/2 x Linear x Exponential + Noise kernel) in the Bedford Ouse, Findhorn and Severn (2012). Lightgreen indicates 2 standard deviations.

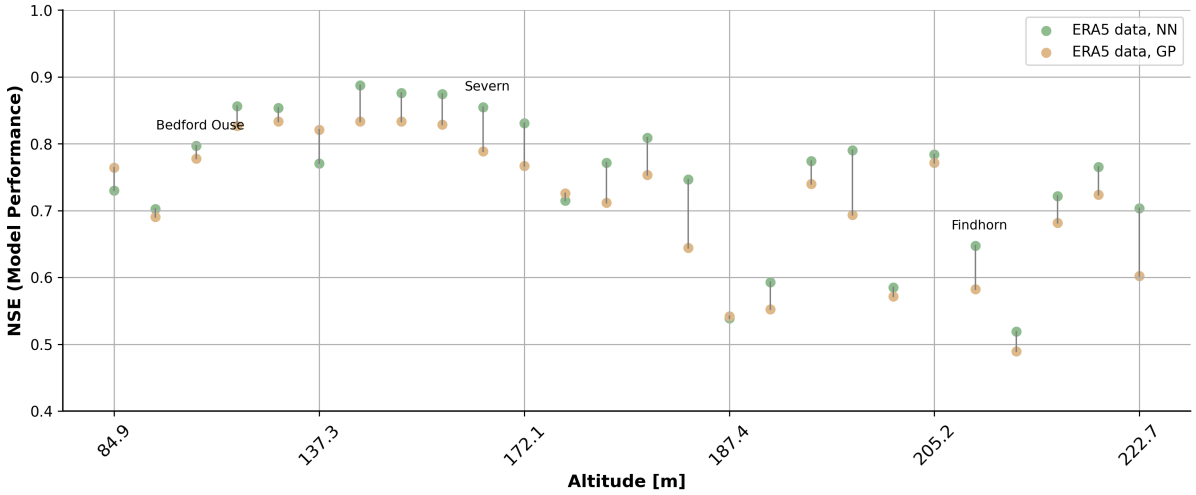


Figure 19: Model performance (NSE) across all 26 catchments with the above described kernel in an SSVGP, compared to a NN applied to the same data.

4 Conclusion

Despite the high correlation between input variables such as precipitation and temperature with streamflow, accurately predicting streamflow remains challenging, particularly for peak values. This challenge is even more pronounced for highland catchments due to various factors, from catchment characteristics to data accuracy. Although no optimal performance has been reached across all catchments, several key insights emerged from this study regarding the performance of models in highland versus lowland catchments.

First, snowmelt has proven crucial for streamflow prediction, particularly in the highlands. Precipitation measurements include snowfall but often overlook the temporal delay needed for snow to melt, especially in higher, colder areas. Various experiments, such as implementing more robust interpolation methods or using higher resolution ERA5-land data, revealed another core issue: a temporal bias in the ERA5 dataset. Using gauged precipitation values from the National River Flow Archive (NRFA) instead of ERA5 reanalysis data demonstrated significant performance improvements (up to 20% NSE improvement), highlighting the critical issue of temporal discrepancies between the datasets.

Models persistently struggle with predicting peak values and these peaks are precisely where ERA5 values significantly differ from NRFA data. A shallow statistical analysis identified the high regime as the most problematic in terms of data similarity. More research is needed to detect potential patterns in the ERA5 data bias and find ways to correct them.

Additionally, the experiments revealed a significant split in model performance between highland and lowland catchments. As uncertainties regarding ERA5 data bias increase in mountainous regions, the quality of input data plays a major role. The nature of the catchment (e.g., small size, high hydrological efficiency) likely contributes, but the experiments could not unambiguously prove this.

Using Gaussian Processes (GPs) provides insights into the relative value of different input variables and reached comparable prediction results. Despite these advantages, GPs face similar predictive challenges as Neural Networks (NNs), re-iterating a persistent geographical discrepancy. Comparing predictions from GPs and NNs could reveal consistent regions of prediction error, possibly overlapping with areas of higher bias in ERA5 data. Additionally, designing GPs to prevent the variance from becoming zero and enhancing their interpretability are necessary steps forward.

In the broader application domain, these findings guide the development of a universal model that effectively handles widely available data products like ERA5. This universality is essential for building robust, adaptable models to address diverse hydrological challenges globally.

References

[1] N.W. Arnell and N.S. Reynard. "The effects of climate change due to global warming on river flows in Great Britain". In: *Journal of Hydrology* 183.3 (1996), pp. 397–424. ISSN: 0022-1694. DOI: https://doi.org/10.1016/0022-1694(95)02950-8. URL: https://www.sciencedirect.com/science/article/pii/0022169495029508.

[2] Thomas G. Huntington. "Evidence for intensification of the global water cycle: Review and synthesis". In: *Journal of Hydrology* 319.1 (2006), pp. 83–95. ISSN: 0022-1694. DOI: https://doi.org/10.1016/j.jhydrol.2005.07.003. URL: https://www.sciencedirect.com/science/article/pii/S0022169405003215.

[3] H. Lee Core Writing Team and J. Romero, eds. *Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Geneva, Switzerland: IPCC, 2023, pp. 35–115. DOI: 10.59327/IPCC/AR6-9789291691647.

[4] Jamie Hannaford. "Climate-driven changes in UK river flows: A review of the evidence". In: *Progress in Physical Geography: Earth and Environment* 39.1 (2015), pp. 29–48. DOI: 10.1177/0309133314536755. eprint: https://doi.org/10.1177/0309133314536755. URL: https://doi.org/10.1177/0309133314536755.

[5] Yukiko Hirabayashi, Mahendran Roobavannan, Sujan Koirala, Lisako Konoshima, Dai Yamazaki, Satoshi Watanabe, Hyungjun Kim, and Shinjiro Kanae. "Global flood risk under climate change". In: *Nature Climate Change* 3 (Sept. 2013), pp. 816–821. DOI: 10.1038/nclimate1911.

[6] P.E Dr. Harry X. Zhang Ph.D. "Climate Change Climate Change and Global Water global water Sustainability climate change global water sustainability". In: Jan. 2012, pp. 2061–2078. DOI: 10.1007/978-1-4419-0851-3_388.

[7] Marzia Ciampittiello, Aldo Marchetto, and Angela Boggero. "Water Resources Management under Climate Change: A Review". In: *Sustainability* 16.9 (2024). ISSN: 2071-1050. URL: https://www.mdpi.com/2071-1050/16/9/3590.

[8] Lindsay C. Stringer, Alisher Mirzabaev, Tor A. Benjaminsen, Rebecca M.B. Harris, Mostafa Jafari, Tabea K. Lissner, Nicola Stevens, and Cristina Tirado-von der Pahlen. "Climate change impacts on water security in global drylands". In: *One Earth* 4.6 (2021), pp. 851–864. ISSN: 2590-3322. DOI: https://doi.org/10.1016/j.oneear.2021.05.010. URL: https://www.sciencedirect.com/science/article/pii/S2590332221002918.

[9] Shan-E-Hyder Soomro, Jiali Guo, Yinghai Li, Zhiqiang Zeng, Abdul Soomro, Sahar Batool, Yanqin Bai, Muhammad Tayyab, Ao Li, Yao Zhen, Kang Rui, Aamir Hameed, Wang Yuanyang, and Caihong Hu. "How does the climate change effect on hydropower potential, freshwater fisheries, and hydrological response of snow on water availability?" In: *Applied Water Science* 14 (Apr. 2024), pp. 1–31. DOI: 10.1007/s13201-023-02070-6.

[10] Matthew LeGrand, James Luce, Robert Metcalfe, and James Buttle. "Development of an Inexpensive Automated Streamflow Monitoring System". In: *Hydrological Processes* 34 (Apr. 2020). DOI: 10.1002/hyp.13783.

[11] Pariva Dobriyal, Ruchi Badola, Chongpi Tuboi, and Syed Ainul Hussain. "A review of methods for monitoring streamflow for sustainable water resource management". In: *Applied Water Science* 7 (Oct. 2016). DOI: 10.1007/s13201-016-0488-y.

[12] Chris Kidd, Andreas Becker, George J. Huffman, Catherine L. Muller, Paul Joe, Gail Skofronick-Jackson, and Dalia B. Kirschbaum. "So, How Much of the Earth's Surface Is Covered by Rain Gauges?" In: *Bulletin of the American Meteorological Society* 98.1 (2017), pp. 69–78. DOI: 10.1175/BAMS-D-14-00283.1. URL: https://journals.ametsoc.org/view/journals/bams/98/1/bams-d-14-00283.1.xml.

[13] J. E. Nash and J. V. Sutcliffe. "River flow forecasting through conceptual models part I — A discussion of principles". In: *Journal of Hydrology* 10.3 (Apr. 1970), pp. 282–290. DOI: 10.1016/0022-1694(70)90255-6.

[14] Bisrat Ayalew Yifru, Kyoung Lim, and Seoro Lee. "Enhancing Streamflow Prediction Physically Consistently Using Process-Based Modeling and Domain Knowledge: A Review". In: *Sustainability* 16 (Feb. 2024), p. 1376. DOI: 10.3390/su16041376.

[15] null null. "Artificial Neural Networks in Hydrology. I: Preliminary Concepts". In: *Journal of Hydrologic Engineering* 5.2 (2000), pp. 115–123. DOI: 10.1061/(ASCE)1084-0699(2000)5:2(115). eprint: https://ascelibrary.org/doi/pdf/10.1061/%28ASCE%291084-0699%282000%295%3A2%28115%29. URL: https://ascelibrary.org/doi/abs/10.1061/%28ASCE%291084-0699%282000%295%3A2%28115%29.

[16] Mohammed Abdul Bari, Urooj Khan, Gnanathikkam Emmanuel Amirthanathan, Mayank Tuteja, and Richard Mark Laugesen. "Simulation of Gauged and Ungauged Streamflow of Coastal Catchments across Australia". In: *Water* 16.4 (2024). ISSN: 2073-4441. DOI: 10.3390/w16040527. URL: https://www.mdpi.com/2073-4441/16/4/527.

[17] Dominique Bourdin, Sean Fleming, and Roland Stull. "Streamflow Modelling: A Primer on Applications, Approaches and Challenges". In: *Atmosphere-Ocean* 50 (Dec. 2012). DOI: 10.1080/07055900.2012.734276.

[18] Robert Edwin Rouse, Doran Khamis, Scott Hosking, Allan McRobie, and Emily Shuckburgh. "Streamflow Prediction Using Artificial Neural Networks & Soil Moisture Proxies". Unpublished manuscript. Unpublished manuscript., Oct. 2023.

[19] A. Sun, Dingbao Wang, and Xianli Xu. "Monthly Streamflow Forecasting Using Gaussian Process Regression". In: *Journal of Hydrology* 511 (Apr. 2014), pp. 72–81. DOI: 10.1016/j.jhydrol.2014.01.023.

[20] F. Kratzert, D. Klotz, G. Shalev, G. Klambauer, S. Hochreiter, and G. Nearing. "Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets". In: *Hydrology and Earth System Sciences* 23.12 (2019), pp. 5089–5110. DOI: 10.5194/hess-23-5089-2019. URL: https://hess.copernicus.org/articles/23/5089/2019/.

[21] UK Centre for Ecology Hydrology. *UK National River Flow Archive Data*. Database. June 2022. URL: https://nrfa.ceh.ac.uk/data.

[22] Arved J. Raudkivi. *Hydrology: An Advanced Introduction to Hydrological Processes and Modelling*. 1st. Print. Oxford: Pergamon Press, 1979.

[23] Manlin Wang, Yu Zhang, Yan Lu, Li Gao, and Leizhi Wang. "Attribution Analysis of Streamflow Changes Based on Large-scale Hydrological Modeling with Uncertainties". In: *Water Resources Management* 37 (Nov. 2022). DOI: 10.1007/s11269-022-03396-7.

[24] H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J-N. Thépaut. *ERA5 hourly data on single levels from 1940 to present*. 2018. DOI: 10.24381/cds.adbb2d47. URL: https://doi.org/10.24381/cds.adbb2d47.

[25] Copernicus Climate Change Service. *ERA5 hourly data on single levels from 1940 to present*. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). 2023. DOI: 10.24381/cds.adbb2d47. URL: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels.

[26] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323.6088 (1986), pp. 533–536. ISSN: 1476-4687.

[27] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. "Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning". In: *CoRR* abs/1702.03118 (2017). arXiv: 1702.03118. URL: http://arxiv.org/abs/1702.03118.

[28] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (2017).

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[30] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006, pp. I–XVIII, 1–248. ISBN: 026218253X.

[31] Elizabeth J. Cross, Timothy J. Rogers, Daniel J. Pitchforth, Samuel J. Gibson, Sikai Zhang, and Matthew R. Jones. "A spectrum of physics-informed Gaussian processes for regression in engineering". In: *Data-Centric Engineering* 5 (2024), e8. DOI: 10.1017/dce.2024.2.

[32] Kenza Tazi, Jihao Andreas Lin, Ross Viljoen, Alex Gardner, ST John, Hong Ge, and Richard E. Turner. *Beyond Intuition, a Framework for Applying GPs to Real-World Data*. 2023. arXiv: 2307.03093 [cs.LG]. URL: https://arxiv.

[33] James Hensman, Alexander G. de G. Matthews, Maurizio Filippone, and Zoubin Ghahramani. *MCMC for Variationally Sparse Gaussian Processes*. 2015. arXiv: 1506.04000 [id='stat.ML' $full_name = 'MachineLearning' is_active = True alt_name = None in_archive = 'stat' is_general = False description = 'Covers machine learning papers (supervised, unsupervised, semi-supervised learning, graphi$

[34] Haibin Yu, Trong Nghia Hoang, Kian Low, and Patrick Jaillet. "Stochastic Variational Inference for Fully Bayesian Sparse Gaussian Process Regression Models". In: (Nov. 2017).

[35] Thomas Pinder and Daniel Dodd. "GPJax: A Gaussian Process Framework in JAX". In: *Journal of Open Source Software* 7 (75 June 2022), p. 4455. DOI: 10.21105/joss.04455. URL: https://github.com/thomaspinder/GPJax.

[36] David Mindham, Keith Beven, and Nick Chappell. "Rainfall–streamflow response times for diverse upland UK micro-basins: quantifying hydrographs to identify the nonlinearity of storm response". In: *Hydrology Research* 54 (Feb. 2023). DOI: 10.2166/nh.2023.115.

[37] Hannah M. Joyce, Jeff Warburton, and Richard J. Hardy. "A catchment scale assessment of patterns and controls of historic 2D river planform adjustment". In: *Geomorphology* 354 (2020), p. 107046. ISSN: 0169-555X. DOI: https://doi.org/10.1016/j.geomorph.2020.107046. URL: https://www.sciencedirect.com/science/article/pii/S0169555X20300167.

[38] Andries Engelbrecht. "Sensitivity Analysis for Selective Learning by Feedforward Neural Networks". In: *Fundam. Inform.* 45 (Aug. 2001), pp. 295–328.

[39] Bart Nijssen, Greg O'Donnell, Alan Hamlet, and Dennis Lettenmaier. "Hydrologic Sensitivity of Global Rivers to Climate Change". In: *Climatic Change* 50 (July 2001), pp. 143–175. DOI: 10.1023/A:1010616428763.

[40] Mingxi Pan, Fang Zhao, Jingyan Ma, Lijuan Zhang, Jinping Qu, Liling Xu, and Yao Li. "Effect of Snow Cover on Spring Soil Moisture Content in Key Agricultural Areas of Northeast China". In: *Sustainability* 14.3 (2022). ISSN: 2071-1050. DOI: 10.3390/su14031527. URL: https://www.mdpi.com/2071-1050/14/3/1527.

[41] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, and J.-N Thépaut. "The ERA5 global reanalysis". In: *Quarterly Journal of the Royal Meteorological Society* 146 (June 2020). DOI: 10.1002/qj.3803.

[42] V. D. J. Keller, M. Tanguy, I. Prosdocimi, J. A. Terry, O. Hitt, S. J. Cole, M. Fry, D. G. Morris, and H. Dixon. "CEH-GEAR: 1 km resolution daily and monthly areal rainfall estimates for the UK for hydrological and other applications". In: *Earth System Science Data* 7.1 (2015), pp. 143–155. DOI: 10.5194/essd-7-143-2015. URL: https://essd.copernicus.org/articles/7/143/2015/.

[43] Zhenyi Zhang, Yidong Lou, Weixing Zhang, Hua Wang, Yaozong Zhou, and Jingna Bai. "Assessment of ERA-Interim and ERA5 reanalysis data on atmospheric corrections for InSAR". In: *International Journal of Applied Earth Observation and Geoinformation* 111 (2022), p. 102822. ISSN: 1569-8432. DOI: https://doi.org/10.1016/j.jag.2022.102822. URL: https://www.sciencedirect.com/science/article/pii/S1569843222000243.

[44] Tao Chen, Liliang Ren, Fei Yuan, Xiaoli Yang, Shanhu Jiang, Tiantian Tang, Yi Liu, Chongxu Zhao, and Liming Zhang. "Comparison of Spatial Interpolation Schemes for Rainfall Data and Application in Hydrological Modeling". In: *Water* 9.5 (2017). ISSN: 2073-4441. DOI: 10.3390/w9050342. URL: https://www.mdpi.com/2073-4441/9/5/342.

[45] Irham Azizan, Samsul Ariffin Abdul Karim, and S Suresh Kumar Raju. "Fitting Rainfall Data by Using Cubic Spline Interpolation". In: *MATEC Web of Conferences* 225 (Jan. 2018), p. 05001. DOI: 10.1051/matecconf/201822505001.

[46] Malcolm N. Mistry, Rochelle Schneider, Pierre Masselot, Dominic Royé, Ben Armstrong, Jan Kyselý, Hans Orru, Francesco Sera, Shilu Tong, Eric Lavigne, Ales Urban, Joana Madureira, David García-León, Dolores Ibarreta, Juan-Carlos Ciscar, Luc Feyen, Evan De Schrijver, Micheline De Sousa Zanotti Stagliorio Coelho, Mathilde Pascal, Aurelio Tobias, Multi-Country Multi-City (MCC) Collaborative Research Network, Barrak Alahmad, Rosana Abrutzky, Paulo Hilario Nascimento Saldiva, Patricia Matus Correa, Nicolás Valdés Orteg, Haidong Kan, Samuel Osorio, Ene Indermitte, Jouni J. K. Jaakkola, Niilo Ryti, Alexandra Schneider, Veronika Huber, Klea Katsouyanni, Antonis Analitis, Alireza Entezari, Fatemeh Mayvaneh, Paola Michelozzi, Francesca de'Donato, Masahiro Hashizume, Yoonhee Kim, Magali Hurtado Diaz, César De La Cruz Valencia, Ala Overcenco, Danny Houthuijs, Caroline Ameling, Shilpa Rao, Xerxes Seposo, Baltazar Nunes, Iulian-Horia Holobaca, Ho Kim, Whanhee Lee, Carmen Íñiguez, Bertil Forsberg, Christofer Åström, Martina S. Ragettli, Yue-Liang Leon Guo, Bing-Yu Chen, Valentina Colistro, Antonella Zanobetti, Joel Schwartz, Tran Ngoc Dang, Do Van Dung, Yuming Guo, Ana M. Vicedo-Cabrera, and Antonio Gasparrini. "Comparison of weather station and climate reanalysis data for modelling temperature-related mortality". en. In: *Scientific Reports* 12.1 (Mar. 2022), p. 5178. ISSN: 2045-2322. DOI: 10.1038/

s41598-022-09049-4. URL: https://www.nature.com/articles/s41598-022-09049-4 (visited on 05/20/2024).

[47] Susan M. Crooks, Alison L. Kay, Helen N. Davies, and Victoria A. Bell. "From Catchment to National Scale Rainfall-Runoff Modelling: Demonstration of a Hydrological Modelling Framework". In: *Hydrology* 1.1 (2014), pp. 63–88. ISSN: 2306-5338. DOI: 10.3390/hydrology1010063. URL: https://www.mdpi.com/2306-5338/1/1/63.

[48] European Centre for Medium-Range Weather Forecasts. *ERA5-Land: data documentation*. 2024.

[49] David Lavers, Adrian Simmons, Freja Vamborg, and Mark Rodwell. "An evaluation of ERA5 precipitation for climate monitoring". In: *Quarterly Journal of the Royal Meteorological Society* 148 (Aug. 2022). DOI: 10.1002/qj.4351.

[50] David Duvenaud. "Automatic model construction with Gaussian processes". In: (Apr. 2015).

[51] Ahmed Elbeltagi, Nasrin Azad, Arfan Arshad, Safwan Mohammed, Ali Mokhtar, Chaitanya Pande, Hadi Ramezani Etedali, Shakeel Ahmad Bhat, Abu Reza Md. Towfiqul Islam, and Jinsong Deng. "Applications of Gaussian process regression for predicting blue water footprint: Case study in Ad Daqahliyah, Egypt". In: *Agricultural Water Management* 255 (2021), p. 107052. ISSN: 0378-3774. DOI: https://doi.org/10.1016/j.agwat.2021.107052. URL: https://www.sciencedirect.com/science/article/pii/S0378377421003176.

[52] Pan Wang, Lu Zhenzhou, Jixiang Hu, and Changcong Zhou. "Sensitivity analysis of the variance contributions with respect to the distribution parameters by the kernel function". In: *Computers Mathematics with Applications* 67 (June 2014). DOI: 10.1016/j.camwa.2014.04.007.

[53] Peter Sollich and Christopher K. I. Williams. "Understanding gaussian process regression using the equivalent kernel". In: *Proceedings of the First International Conference on Deterministic and Statistical Methods in Machine Learning*. Sheffield, UK: Springer-Verlag, 2004, pp. 211–228. ISBN: 3540290737. DOI: 10.1007/11559887_13. URL: https://doi.org/10.1007/11559887_13.

[54] Andrew Mchutchon and Carl Rasmussen. "Gaussian Process Training with Input Noise". In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger. Vol. 24. Curran Associates, Inc., 2011. URL: https://proceedings.neurips.cc/paper_files/paper/2011/file/a8e864d04c95572d1aece099af852d0a-Paper.pdf.

[55] C. Fyfe, Tzai-Der Wang, and Shang Chuang. "Comparing Gaussian Processes and Artificial Neural Networks for Forecasting". In: Jan. 2006. DOI: 10.2991/jcis.2006.7.

[56] Lu Cheng, Siddharth Ramchandran, Tommi Vatanen, Niina Lietzén, Riitta Lahesmaa, Aki Vehtari, and Harri Lähdesmäki. "An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data". In: *Nature Communications* 10 (Apr. 2019), p. 1798. DOI: 10.1038/s41467-019-09785-8.

[57] Luis Damiano, Margaret Johnson, Joaquim Teixeira, Max D. Morris, and Jarad Niemi. *Automatic Dynamic Relevance Determination for Gaussian process regression with high-dimensional functional inputs*. 2022. arXiv: 2209.00044 [stat.ME].

[58] Tadej Krivec, Juš Kocijan, Matija Perne, Boštjan Grašic, Marija Zlata Božnar, and Primož Mlakar. "Data-driven method for the improving forecasts of local weather dynamics". In: *Engineering Applications of Artificial Intelligence* 105 (2021), p. 104423. ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2021.104423. URL: https://www.sciencedirect.com/science/article/pii/S0952197621002712.

[59] Tayeb Boulmaiz, Kacem Gairaa, Mawloud Guermoui, and Boutaghane Hamouda. "Streamflow forecasting using Gaussian Process Regression Methodology". In: Nov. 2019.

[60] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

[61] Osval Montesinos-López, Abelardo Montesinos, and Jose Crossa. "Fundamentals of Artificial Neural Networks and Deep Learning". In: Jan. 2022, pp. 379–425. ISBN: 978-3-030-89009-4. DOI: 10.1007/978-3-030-89010-0_10.

[62] Bertil Matérn. *Spatial Variation*. Vol. 36. Lecture Notes in Statistics. New York, NY: Springer New York, 1986. ISBN: 978-0-387-96365-5 978-1-4615-7892-5. DOI: 10.1007/978-1-4615-7892-5. URL: http://link.springer.com/10.1007/978-1-4615-7892-5 (visited on 01/15/2023).

5 Appendix

5.1 Notes on ERA5 Outputs

Data Type	CDS Entry	Aggregation	Data Source
precipitation	total_precipitation	daily aggregation	reanalysis-era5-single-levels
temperature	temperature	value at 12pm	reanalysis-era5-single-levels
resultant windspeed	u_component_of_the_wind	value at 12pm	reanalysis-era5-single-levels
	v_component_of_the_wind	value at 12pm	reanalysis-era5-single-levels
humidity	relative_humidity	value at 12pm	reanalysis-era5-single-levels
snow	snow_melt	daily aggregation	reanalysis-era5-single-levels
soil moisture	volumetric_soil_water_layer_1	value at 12pm	reanalysis-era5-single-levels
	volumetric_soil_water_layer_2		reanalysis-era5-single-levels
	volumetric_soil_water_layer_3		reanalysis-era5-single-levels
	volumetric_soil_water_layer_4		reanalysis-era5-single-levels
precipitation	total_precipitation	daily aggregation	reanalysis-era5-land
snow	snow_melt	daily aggregation	reanalysis-era5-land

Table 6: All ERA5 inputs used in this research. All fields were extracted from 1/1/1980 to 31/12/2020 on the provided 0.25° (ERA5) or 0.10° (ERA5-land) latitude-longitude grid. Aggregation methods are included.

5.2 Catchment Overview

Location	**Altitude (m)**
Dee at Polhollick	812.1
Spey at Kinrara	811.2
Glass at Kerrow Wood	795.6
Glass at Fasnakyle	761.0
Findhorn at Shenachie	663.6
Broom at Inverbroom	528.3
Devon at Glenochil	511.0
Ure at Westwick	464.1
Swale at Crakehill	440.0
Wyre at Scorton Weir	440.0
Bogie at Redcraig	427.2
Coquet at Morwick	420.0
Derwent at St Mary's Bridge	414.0
Derwent at Church Wilne	404.4
Exe at Pixton	398.4
Teme at Tenbury	383.7
Severn at Haw Bridge	303.5
Trent at Colwick	273.7
Trent at North Muskham	260.2
Thames at Eynsham	222.7
Kennet at Theale	205.2
Thames at Royal Windsor Park	187.4
Avon at Bathford	172.1
Bedford Ouse at Roxton	137.3
Ely Ouse at Denver Complex	84.9
Wensum at Costessey Mill	71.1

Table 7: Catchment names, locations and their corresponding altitudes, ordered by descending altitude.

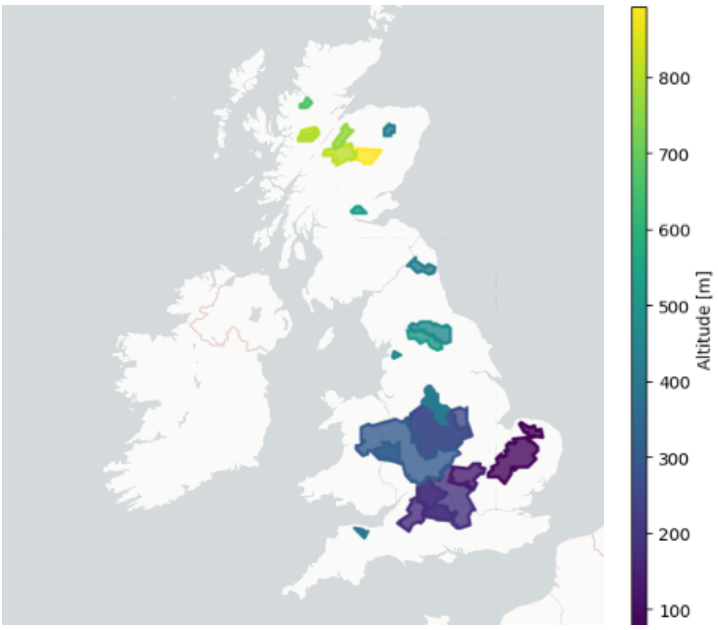


Figure 20: Overview of the 26 studied catchments across the UK. The colorbar indicates the altitude of the catchment.

5.3 Data Interpolation

5.3.1 Bilinear interpolation

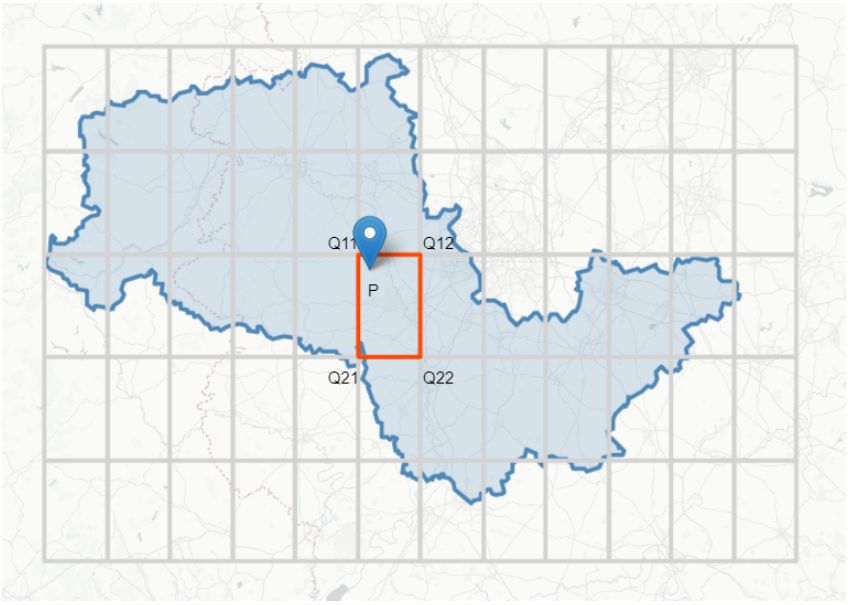


Figure 21: Bedford Ouse catchment over the 0.25° by 0.25° ERA5 grid. The red box indicates the four points used for bilinear interpolation at the centroid.

The bilinear interpolation formula for a point $P(x, y)$ given four known points Q_{11}, Q_{12}, Q_{21}, and Q_{22} is given by:

$$P(x,y) = \frac{(x_2 - x)(y_2 - y)}{(x_2 - x_1)(y_2 - y_1)} Q_{11} + \frac{(x - x_1)(y_2 - y)}{(x_2 - x_1)(y_2 - y_1)} Q_{12} + \frac{(x_2 - x)(y - y_1)}{(x_2 - x_1)(y_2 - y_1)} Q_{21} + \frac{(x - x_1)(y - y_1)}{(x_2 - x_1)(y_2 - y_1)} Q_{22} \tag{9}$$

where:

$$Q_{11} \text{ is the value at } (x_1, y_1),$$
$$Q_{12} \text{ is the value at } (x_2, y_1),$$
$$Q_{21} \text{ is the value at } (x_1, y_2),$$
$$Q_{22} \text{ is the value at } (x_2, y_2).$$

5.3.2 Surface interpolation

Surface interpolation within this research refers to bicubic interpolation, a method used to interpolate data points on a two-dimensional grid (dimensions i, j) to produce a smooth surface that fits the data. Given a set of points with coordinates (x_i, y_i, z_i), where x_i and y_i are the grid points and z_i is the value (e.g., precipitation) at those points, bicubic interpolation estimates the value at any point (x, y) within the grid [44]. The bicubic interpolation formula can be written as:

$$f_p(x, y) = \sum_{i=0}^{3} \sum_{j=0}^{3} a_{ij} (x - x_0)^i (y - y_0)^j \tag{10}$$

where:

- a_{ij} are the coefficients determined by the function values and the derivatives at the grid points.
- (x_0, y_0) is the reference grid point from which the interpolation is performed.

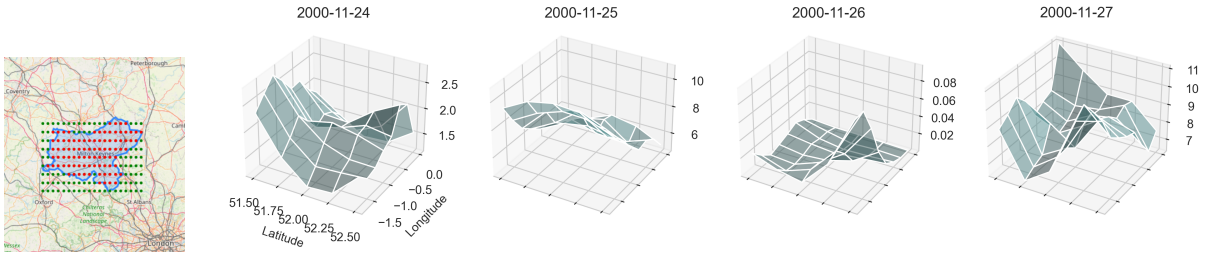
The coefficients a_{ij} can be calculated using the function values and the partial derivatives at the grid points. This involves solving a system of linear equations formed by the function values, the first-order partial derivatives, and the mixed second-order partial derivatives at the grid points.

Once the interpolated function $f_p(x, y)$ is established, this function can be integrated over the two-dimensional surface S, representing the catchment. The catchment surface S will always have an irregular shape, which is not easily integrated. To avoid the complexity of parametrization and Jacobians, a grid-based numerical method approximates the integral of the interpolated function $f_p(x, y)$ over surface S:

$$\iint_S f_p(x, y)\, dA \approx \sum_{k=1}^{N} f_p(x_k, y_k) \Delta A_k \tag{11}$$

Where:

- (x_k, y_k) are the grid points within the surface S.
- ΔA_k is the area element associated with each grid point (x_k, y_k).
- N is the total number of grid points within the surface S.



(a) ERA5 Grid. (b) Interpolated precipitation surfaces over 4 days.

Figure 22: The ERA5 0.25°x 0.25° grid resolution for the Bedford Ouse catchment, with the interpolated precipitation surfaces over this grid for 4 days in November 2011.

First the surface S needs to be divided into a grid of small, equal-sized elements (empirically decided on 0.05°). Formula 10 returns the value of the function at the center of each grid element. Multiplying

each value by the area of the corresponding grid element and summing those together results in the approximation of the integral. To obtain the interpolated precipitation value per catchment the last step can be omitted, as the catchment surface is independent of the training inputs. Instead of integrating, all the values over the catchment together divided by the amount of points also gives a fair representation of the precipitation distribution.

5.4 Training Details

5.4.1 Neural Networks

The Neural Network used is a Multilayer Perceptron (MLP), which is a feedforward neural network that is composed of multiple layers of neurons [26]. As shown in table 8, each layer in an MLP consists of a linear transformation followed by a non-linear activation function. The activation function is the SiLu (Sigmoid Linear Unit) and is defined as:

$$\text{SiLU}(x) = x \cdot \frac{1}{1 + e^{-x}} \tag{12}$$

The SiLu activation function combines the input with its sigmoid, allowing a smooth and continuous transition [27].

Layer	Description	Output Dimensions
input layer	input	# input features
hidden layer	Fully Connected Layer	64
	SiLU Activation	64
	Dropout Layer (0.2)	64
hidden layer	Fully Connected Layer	16
	SiLU Activation	16
	Dropout Layer (0.2)	16
output layer	Fully Connected Layer	1

Table 8: Structure of the neural network.

The network as presented in 8 contains a set of parameters, that should optimised in order to predict streamflow values as possible to the targets (measured streamflow values). This optimisation process is the NN training [26]. The model performance is evaluated using the RMSE (Root Mean Squared Error, see 2.5), measuring the difference between predicted and actual values. The goal is to minimize this loss, using an optimizer (in this case AdamW). This optimizer updates the model's weight, with a certain learning rate (step size), in this study set at 0.001 [28]. Backpropagation is used to calculate the gradients of the loss function with respect to the model's weights, which indicate the direction and magnitude of adjustments needed to minimize the loss [60].

To prevent overfitting, dropout and early stopping are implemented. Dropout randomly deactivates a fraction of the neurons in each training iteration, to learn more robust features. Early stopping is implemented through a validation set, if no improvement is observed over a period of 10 epochs, the training is halted. Total maximal training time is 9000 epochs [61].

5.4.2 Gaussian Processes

A Gaussian Process (GP) is a probabilistic model used to describe a distribution over functions [32]. It is defined by its mean function $\mu(x)$ and a covariance (kernel) function $k(x, x')$:

$$\mu(x) = \mathbb{E}[f(x)] \tag{13}$$

$$k(x, x') = \mathbb{E}[(f(x) - \mu(x)(f(x') - \mu(x')] \tag{14}$$

$f(x)$ represents the value of the underlying function at the input point x. The mean function $\mu(x)$ gives the expected value of $f(x)$, providing an average prediction. For simplicity, the mean function is assumed to be zero, $\mu(x) = 0$. The covariance function $k(x, x')$ or kernel defines the relationship between the function values at different points x and x', determining how much the value of $f(x)$ at one point can tell us about the value of $f(x')$ at another point [62]. The kernel design is crucial because it defines how data points are correlated with each other [30]. Based on the input data, the kernel components used for streamflow prediction in this research are the Matérn 3/2, exponential, linear and white noise kernel:

$$k_{\text{Matérn-3/2}}(x, x') = \sigma^2 \left(1 + \frac{\sqrt{3}\|x - x'\|}{l}\right) \exp\left(-\frac{\sqrt{3}\|x - x'\|}{l}\right) \tag{15}$$

$$k_{\text{Exponential}}(x, x') = \sigma^2 \exp\left(-\frac{\|x - x'\|}{l}\right) \tag{16}$$

$$k_{\text{Linear}}(x, x') = \sigma^2(x \cdot x' + c) \tag{17}$$

$$k_{\text{White Noise}}(x, x') = \sigma^2 \delta(x - x') \tag{18}$$

where:

- σ^2 is the variance,
- l is the lengthscale,
- $\|x - x'\|$ is the Euclidean distance between points x and x'
- $\delta(x - x')$ is the Kronecker delta function, which is 1 if $x = x'$ and 0 otherwise.

The combination of the these kernel components results in the prior, from which random samples can be drawn, as illustrated in figure 23.

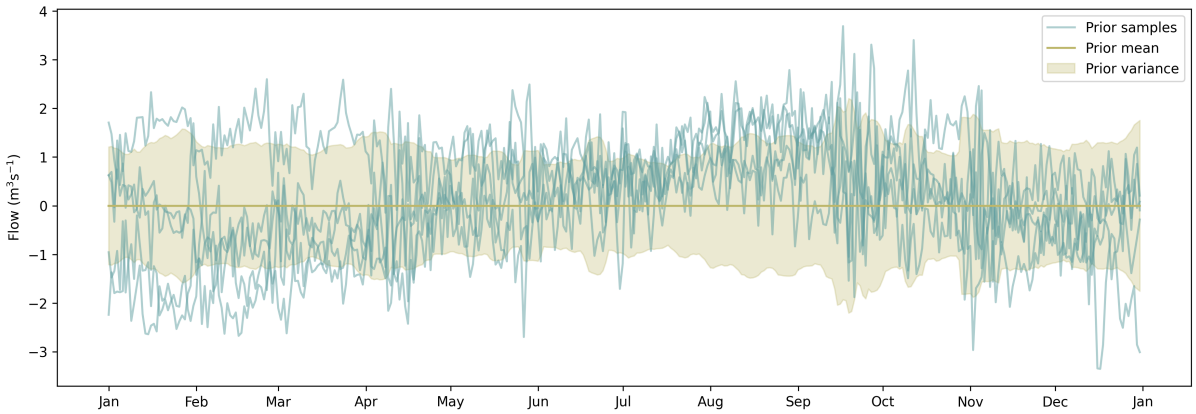


Figure 23: Random samples drawn from the prior as described in formula 8. Input to the samples are the meteorological input variables, resulting in initial suggestions for the streamflow output.

.

Given training data $\{X, y\}$, the prior belief can be updated to form the posterior distribution, through Bayesian inference [30]. The mean of the posterior distribution (denoted as μ_*) represents the GP's best estimate of the function value at new input points X_*, while the uncertainty associated with these predictions is quantified through the predictive variance σ_*^2 [32], allowing to quantify the confidence in the estimates.

$$\mu_*(x_*) = k_{*n}^\top (K + \sigma_n^2 I)^{-1}(y - \mu) + \mu(x_*) \tag{19}$$

$$\sigma_*^2(x_*) = k_{**} - k_{*n}^\top (K + \sigma_n^2 I)^{-1} k_{n*} \tag{20}$$

where:

- $k_{*n} = [k(x_*, x_1), \ldots, k(x_*, x_n)]^\top$
- $k_{**} = k(x_*, x_*)$ is the covariance matrix of the new inputs
- $K = k(x, x')$ is the covariance matrix of the training inputs
- σ_n^2 is the variance of the Gaussian noise
- I is the identity matrix

For traditional GP's, the computational complexity is dominated by the need to invert the covariance matrix and compute the determinant of the covariance matrix, both of which involve operations on an $n \times n$ matrix, where n is the number of training data points [33]. This results in a memory complexity that scales cubically with the number of data points, which makes them impractical for large datasets [32]. To mitigate this, Sparse Stochastic Variational Gaussian Processes (SSVGPs) are used. SSVGPs use a subset of the original data, the so-called inducing points, to approximate the full GP. These points are treated as additional parameters to be optimized. The optimisation (through stochastic gradient descent (SGD)) involves the introduction of a variational distribution, that should be as close as possible to the true posterior. The quality of this approximation is measured using the Evidence Lower Bound (ELBO) [34]. This ELBO is maximised, while making sure that the variational distribution fits the observed data (maximising the likelihood) and ensuring that the variational distribution remains close to the prior distribution (to avoid overfitting) [33].

$$\text{ELBO} = \mathbb{E}_{q(f)}[\log p(y|f)] - \text{KL}(q(f) \| p(f)) \tag{21}$$

where:

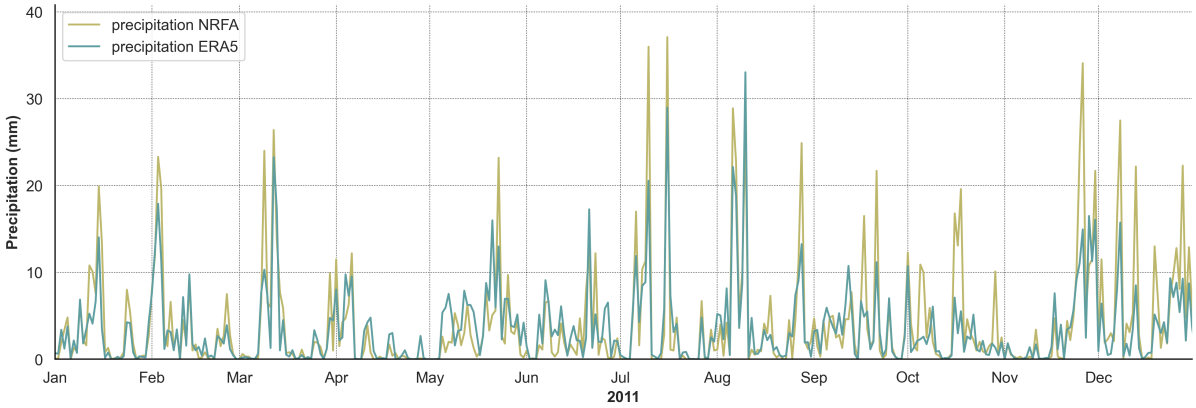
- $\mathbb{E}_{q(f)}[\log p(y|f)]$ is the expected log-likelihood of the observed data under the variational distribution,
- $\text{KL}(q(f)||p(f))$ is the Kullback-Leibler (KL) divergence between the variational distribution $q(f)$ and the prior distribution $p(f)$.

The final posterior distribution provides both the predictive mean and uncertainty, similar to traditional GPs, but with significantly reduced computational requirement. The selected set of m inducing points, m is typically much smaller than n, reducing the number of data points the model interacts with [33].

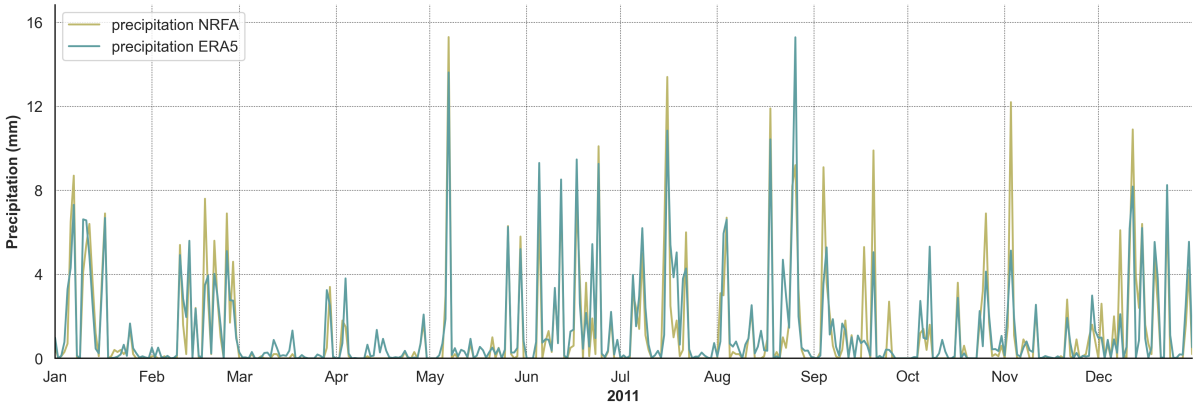
	GPs	**SSVGPs)**
Training Complexity	$\mathcal{O}(n^3)$	$\mathcal{O}(nm^2)$
Memory Complexity	$\mathcal{O}(n^2)$	$\mathcal{O}(nm + m^2)$

Table 9: Computational complexities of Gaussian Processes (GPs) and Sparse Variational Gaussian Processes (SVGPs).

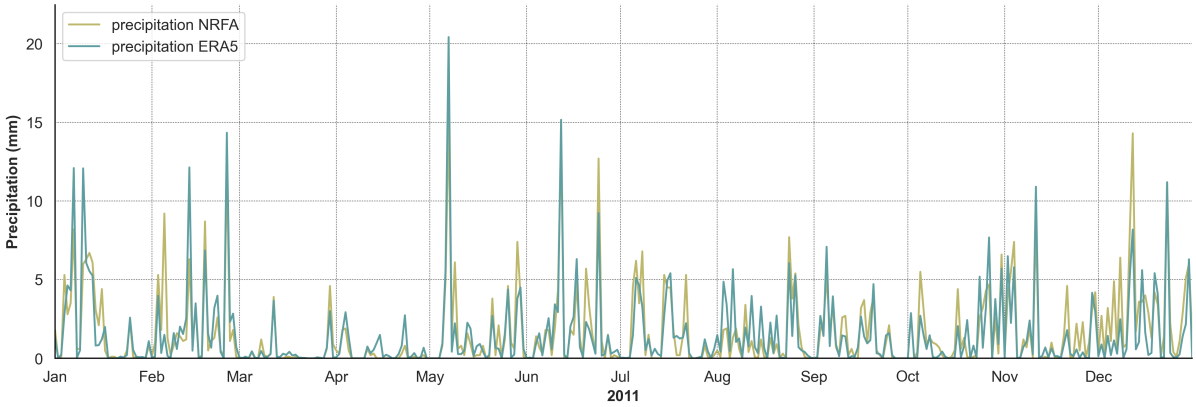
5.5 Precipitation data bias



(a) Findhorn at Shenachie



(b) Bedford Ouse at Roxton



(c) Severn at Haw Bridge

Figure 24: Precipitation values as recorded by ERA5 (reanalysis) versus NRFA (measurements) in 2011 across the three focus catchments.