Utrecht University

Faculty of Geosciences

Earth Surface and Water master thesis

Enhancing Global Streamflow Predictions: Integrating Remote Sensing Data into a Hybrid PCR-GLOBWB and Random Forest Modeling Approach

First examiner:

Prof. dr. Derek Karssenberg

Candidate:

Sümeyye Büşra Işık

Second examiner:

Dr. Edwin Sutanudjaja

Third examiner:

Oriol Pomarol Moya MSc

March 31, 2024

Abstract

Streamflow predictions are essential for effective water management, enabling assessment of water availability, maintenance of agricultural practices, flood mitigation, and the overall impact on society. Traditional hydrological models have a difficult time comprehending some of the complex behavior of the hydrological cycle, leading them to have less accurate streamflow predictions. A hybrid modeling approach that couples the PCR-GLOBWB global hydrological model with the Random Forest machine learning algorithm was developed to enhance streamflow predictions. This study presents an innovative method for improving hydrological simulations by integrating satellite-derived precipitation and evaporation data into this hybrid modeling setup. The research explores the impact of this integration across different global contexts and uncovers the varying efficacy of satellite data in enhancing model accuracy, especially in regions with limited data. Through a comparative analysis of model performances using both global and local training datasets, the study emphasizes the critical importance of using satellite-based data and the strategic use of localized data for optimal predictions. The findings of this study show that model performance did not improve with the integration of satellite-based evaporation and precipitation globally. However, it suggests that satellite data integration offers significant benefits in certain contexts, even though its overall impact depends on the specific hydrological and geographical characteristics of the target region. This research provides valuable insights into the potential use of satellite data to enhance the accuracy and reliability of hydrological predictions, creating opportunities for more informed water resource management strategies amid global environmental changes.

Contents

Appendix

1. Introduction

Streamflow plays an important role in the hydrological cycle. It has been significantly impacted by climate change, leading to more frequent occurrences of extreme weather events like droughts and floods [1]. Streamflow is also essential in water management as it significantly impacts various aspects, such as the availability of water resources, agricultural practices, and accurate flood forecasting. Water resource management could be defined as preparing, creating, supplying, and controlling the optimal use of water resources [2]. From a water resources management perspective, identifying trends and variability in streamflow is critical for planning purposes [3]. Therefore, accurate predictions of streamflow are essential for effective water resource management. Hydrological models have been carefully developed to simulate the behavior of the hydrological cycle and predict its response under changing climatic conditions. Over time, various advanced statistical and computational modeling techniques have been introduced for more precise streamflow simulations and forecasting. Operational hydrological forecasting systems are essential in water resources management and preparedness against extreme events [4]. However, it is difficult to predict the changes in future streamflows as it involves a physical process that depends upon more than one variable, such as precipitation, evapotranspiration, topography, and human activities [5]. Hydrological models can be classified into physically based, conceptual, and/or data-driven models. Each type has advantages and limitations in simulating the complex processes of the hydrological cycle.

Physically-based models are developed based on the understanding of the runoff generation processes, transport in channels, and mathematical formulations of these physical processes [6]. They usually do not attempt to consider the stochastic nature of the underlying hydrologic system [5]. Conceptual models simplify the representation of the hydrological system by combining various parameters and processes into a single equation or set of equations. These models use storage elements as their main components, filled by inputs like rainfall, infiltration, or percolation and emptied through evapotranspiration, runoff, and drainage [7]. The increase in global temperatures is causing shifts in precipitation patterns, melting of glaciers and permafrost, and changes in the frequency and intensity of extreme weather events, all of which are altering river streamflow, flood patterns, flow duration curves, and low-flow periods and posing threats to human societies and ecosystems [8]. Therefore, it is crucial for a hydrological model to accurately simulate rainfall-runoff under changing climatic conditions. However, according to Dakhlaoui, Merz et al. [9] suggested that the climate dependence of model parameters seriously questions the validity of conceptual hydrological models under climate change since model parameters are supposed to represent the physical catchments characteristics without being influenced by climate conditions [10].

One of the most recent advancements in hydrological modeling is the introduction of data-driven models. Unlike physically-based or conceptual models, data-driven models rely on empirical relationships derived from historical data to predict streamflow behavior. These models consist of machine learning algorithms and/or statistical techniques that allow the analysis of large datasets and the identification of patterns in the data. Data-driven models are constrained by the quality of the input data, and the preprocessing techniques used can significantly impact their performance. Another limitation of these types of models that they cannot adapt to process changes brought by physical modifications in the catchment area. Consequently, it is doubtful whether they can be applied to scenarios, as the relationships derived from data may prove invalid in future climates [11]. Another common challenge of data-driven models is overfitting, which means that noise within the data could negatively impact the models predictive performance when handling new data due to the lack of understanding of the physical hydrological processes [12].

During the last three decades, hydrologists, water managers, and forecasters have worked hard to improve the forecasting accuracy of data-driven models by adopting several tools and techniques with a wide variety of computational algorithms and advancing the subject of streamflow modeling [5]. Given the strengths and weaknesses of each type of hydrological model, a hybrid modeling approach that integrates data-driven techniques with physically-based or conceptual models may offer a promising solution. General hydrological models remain essential tools for accurate streamflow predictions. However, their complex nature and limited ability to act upon insufficient data, oversimplify certain parameters, and lack of adaptation to climatic changes have prompted the implementation of alternative methods to enhance performance.

Hybrid model combines global performance with superior local adaptivity, surpassing physically based models by replacing complex physical processes and integrating diverse datasets through highly data-adaptive neural network parameterization [13]. A hybrid modeling approach integrates a statistical learning algorithm with a physically-based hydrological model, combining the ability to handle large datasets and identify complex patterns with physical process representation. This approach can capture the stochastic nature of the hydrological system, adapt to process changes caused by modifications in the catchment area, and improve forecast accuracy. Artificial Intelligence (AI) offers many popular data-driven models which have been used extensively in the past couple of decades in different aspects of hydrology, including stream flow forecasting, evapotranspiration estimation, solar radiation modeling and rainfall-runoff modelling [14]. The use of machine learning algorithms, such as artificial neural networks (ANNs) and long short-term memory (LSTM) models, can help to identify the most relevant predictors and generate an ensemble of different climate model predictions [15]. While some attractive properties of random forests are also shared by other data-driven methods, their selection is driven mostly by their increased predictive performance, their capability to capture non-linear dependencies and interactions of variables, as well as their speed, parsimonious parameterization, ease of use, and ability to handle big datasets [16]. By adopting a hybrid modeling approach, water managers, hydrologists, and forecasters can benefit from improved streamflow predictions, enhanced adaptivity to changing climatic conditions, and the integration of various datasets. In conclusion, hybrid modeling presents a promising solution for addressing the challenges associated with individual hydrological modeling techniques and advancing the accuracy and reliability of streamflow predictions.

Shen et al.[17] developed a model that utilizes the Random Forest method to improve PCR-GLOBWB daily streamflow predictions of the PCR-GLOBWB global hydrological model[18]. Three gauging stations along the Rhine River were selected to investigate how catchment characteristics influence error correction. Meteorological data and state variables of PCR-GLOBWB were used to estimate its prediction errors. Then, another error estimation was made using the Random Forest method. RF error estimation has been added to PCR-GLOBWB for final corrected predictions. The study has shown that the RF method improved the performance of the PCR-GLOBWB model. The equal performance of calibrated and uncalibrated models indicated that RF can be used for successful error corrections without going through a complex model calibration process.

Work of Shen et al. [17] was extended to a global scale by Magni et al. [19] in order to explore the potential of using statistical learning methodology as a proxy for improving streamflow predictions in ungauged basins. More predictors were introduced into the model, such as static catchment attributes, meteorological input, hydrological state variables, and simulated runoff from the global hydrological model PCR-GLOBWB. The response variable for the RF model was changed to streamflow observation instead of error estimation. They achieved significant improvements for most stations.

Collot d'Escury [20] applied some improvements to Magni et al.'s hybrid modeling by introducing satellite data as an input for the post-processor. Liquid Water Equivalent (LWE), Snow Cover Fraction (SCF), and Soil Moisture (SM) from remote sensing observations were implemented in the hybrid model work of Magni et al. along with meteorological variables, streamflow observations, and hydrological state variables from PCR-GLOBWB. Different input combinations were tested to see whether the complete removal of PCR-GLOBWB variables is possible. The study showed that using specified remote sensing data did not improve global runs; however, complete or partial removal of variables can still maintain the model's high performance. Collot d'Escury's study emphasized the importance of using satellite-based data as an alternative to model variables. Using satellite data as an alternative to model variables showed promising results in improving the precision of streamflow predictions while reducing the number of predictors and computational time required.

Considering the success of Collot d'Escury, other satellite products might be valuable for improving streamflow predictions. Exploring alternative approaches, such as utilizing satellite-based evaporation and precipitation estimates as inputs, could further enhance the precision of streamflow predictions. Precipitation and evaporation are the two key components of the global water cycle [21]. With the advancement of remote sensing technology, satellite-based precipitation products have become an effective supplement for measured data [22]. Experiments by Alfieri et al. showed that satellite-derived GLEAM evaporation data led to a 2% improvement over baseline runs driven by high-quality ground-based datasets [4]. The GLEAM data offer spatial coherence, global validity, satellite-derived observations, a minimalistic approach, validation against ground measurements, and flexibility in application, making them valuable for studying global land-surface evaporation dynamics and their implications for water and climate assessments [21]. IMERG precipitation data offer global coverage, high spatial and temporal resolution, integration of multiple data sources, quality indices for reliability assessment, differentiation of precipitation phases, and consistent processing [23]. By integrating these alternative data sources into the modeling framework, it might be possible to enhance the accuracy of streamflow predictions while reducing the computational burden associated with a large number of predictors. Building on previous research by Magni et al.[19] and using data from two satellites, the study aims to investigate streamflow predictions and assess the impact of satellite-based data on the performance of a hybrid hydrological model. By examining different dataset configurations at both global and local scales, this study seeks to systematically analyze the effects of incorporating satellite-based data in the modeling process. Consequently, it will provide insights into the following questions

- How does the integration of satellite-based precipitation and evaporation affect the performance of the hybrid hydrological model in predicting streamflow?

For a more comprehensive understanding of the main research question, this study will address the following sub-questions:

- Is it possible to completely or partially remove PCR-GLOBWB products as an input to the model and still achieve satisfactory results?
- Does a model trained on local data outperform a model trained on global data for the same local area?
- How does the model perform in regions with varying availability of ground-based observations?
- How does the model perform in regions with different climatic conditions with similar data availability?

The results of this research aim to contribute to the understanding of the potential benefits and limitations of incorporating satellite products in a hybrid hydrological model.

The thesis is organized into several key sections. The second (2) section outlines the sources of data used in the study. Section 3 details the approach taken to integrate satellite data into the hybrid hydrological model. This section also provides specific configurations and setups used for the modeling process. Section 4 presents the findings of the study, including performance comparisons with and without satellite inputs at both global and local scales. Discussions are presented in section 5 where implications of results are evaluated and compared with existing literature to provide a deeper understanding of their significance in hydrological modeling using satellite products. Finally, section 6 contains conclusions summarizing key findings along with their implications, offering insights into potential benefits as well as limitations of incorporating satellite products in a hybrid hydrological model.

2. Data and Methods

2.1 Data

The study utilizes data obtained from previous research and two distinct satellite products. Information sourced from Magni et al. is available at Zenodo [24], encompassing streamflow measurements, meteorological factors, and hydrological state variables derived from the PCR-GLOBWB model between 1979 and 2019. Data for streamflow can be obtained by visiting the Global Runoff Data Centre (GRDC) website. Stations were selected based on a minimum upstream area of 10,000 km^2 and the availability of recorded data for at least one month between 2000 and 2019; a total of 1342 stations have available data, as shown in Figure 2.1.

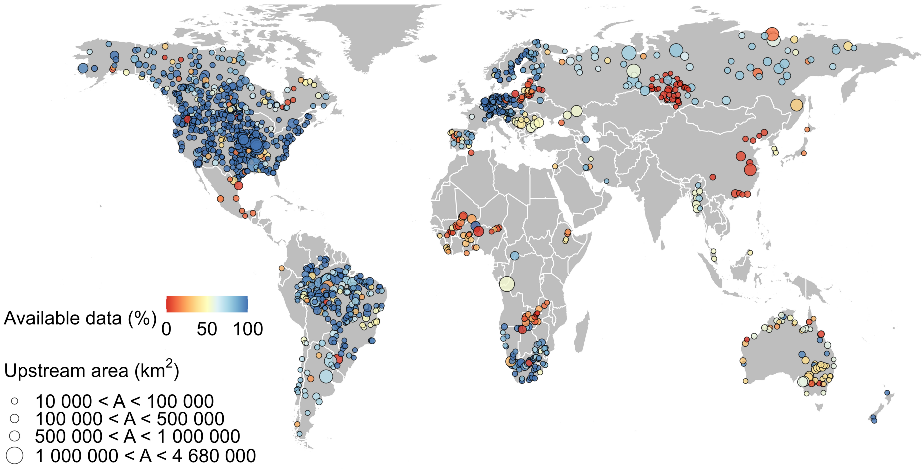


Figure 2.1: Availability (%) of monthly river discharge data spanning from 2000 to 2019 for GRDC stations globally. Circle sizes represent catchment areas, while the color scale indicates the percentage of available data.

2.1.1 GLEAM evaporation data

The Global Land Evaporation Amsterdam Model (GLEAM) is a satellite-based model that helps estimate land evaporation worldwide. The model represents a new approach that combines a wide range of currently existing satellite-sensor products to estimate reliable fields of daily global evaporation at a 0.25-degree spatial resolution [21]. It works by integrating satellite-observed geophysical variables, such as soil moisture, vegetation optical depth, and snow-water equivalent, along with reanalysis of air temperature and radiation and a multi-source precipitation product which drives the model and estimates terrestrial evaporation and root-zone soil moisture [25].

Actual evaporation data from the model datasets for 2000 and 2019 have been selected for this purpose.

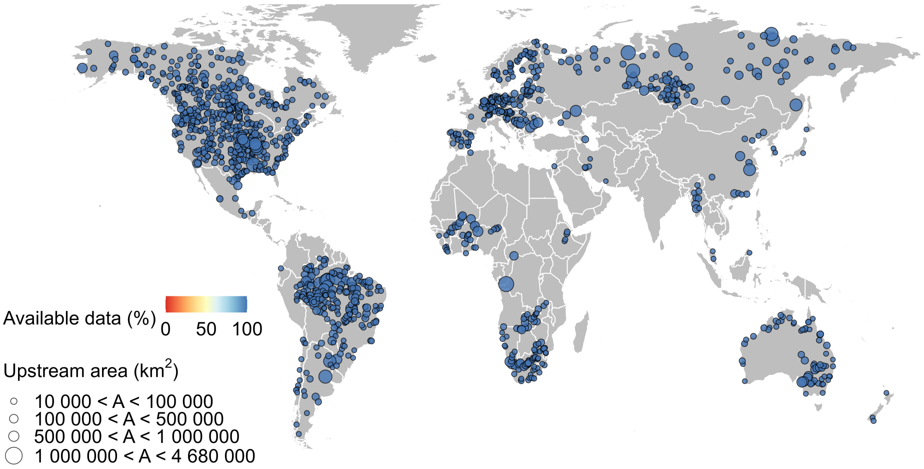


Figure 2.2: Availability (%) of monthly satellite-based evaporation data spanning from 2000 to 2019 for GRDC stations globally. Circle sizes represent catchment areas, while the color scale indicates the percentage of available data.

2.1.2 IMERG precipitation data

The Integrated Multi-satellite Retrievals for GPM (IMERG) is a precipitation dataset that combines information from multiple satellite sensors to provide global precipitation estimates. The GPM satellite constellation consists of one core observatory satellite and about ten partner satellites, equipped with the latest Dual-frequency Precipitation Radar, conical-scanning multichannel GPM Microwave Imager, and many other advanced instruments [26]. The IMERG mission utilizes intercalibrated estimates from satellite passive microwave sensors, microwave-calibrated infrared satellite estimates, and surface precipitation gauge analyses processed, intercalibrated, and combined to produce high-quality, half-hourly gridded datasets for global precipitation measurement and research [23]. The high-quality precipitation estimates and long-term coverage of IMERG are expected to provide insights into future hydro-meteorological processes and climatological studies. IMERG provides precipitation data with a spatial resolution of 0.5 degrees. Coverage is provided for latitudes between 60°N and 60°S, with partial coverage extending to 90° [27], resulting in low coverage in areas close to the polar regions as seen in Figure 2.3

Satellite-derived precipitation data for the years 2000 and 2019 were chosen for this objective.

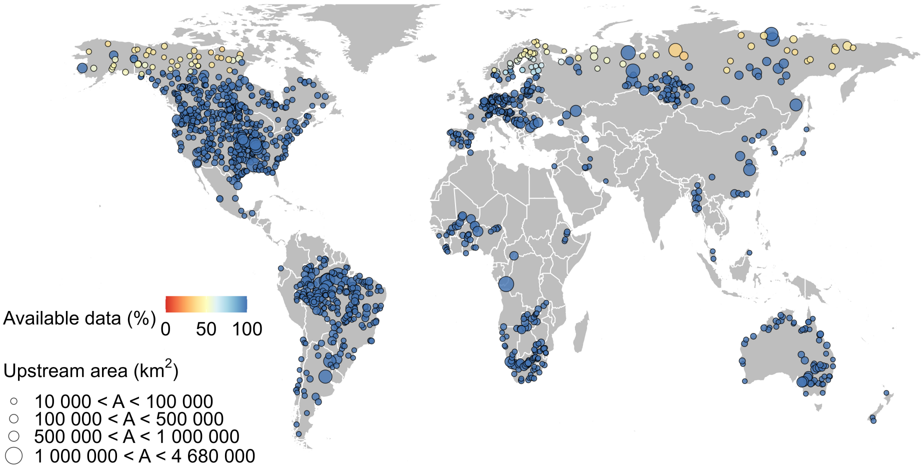


Figure 2.3: Availability (%) of monthly satellite-based precipitation data spanning from 2000 to 2019 for GRDC stations globally. Circle sizes represent catchment areas, while the color scale indicates the percentage of available data.

2.1.3 Data pre-processing

The dataset provided by Magni et al. [19] has already undergone pre-processing, so there is no need for additional manipulation. All the satellite products were rescaled to 0.5° spatial resolution to ensure consistency in the data, using the Climate Data Operators [28] to process the data. Normalization of GLEAM evaporation and IMERG precipitation datasets have been performed in the PCRaster Python framework [29]. Normalization as a pre-processing step was performed using the formula provided below.

$$x_{norm_i} = \frac{x_i - \mu}{\sigma} \tag{2.1}$$

where:

- x_{norm_i} is the normalized value of a data point.
- x_i is the original value of the data point.
- μ is the mean,
- σ is the standard deviation.

After normalizing the data, satellite data was extracted from netCDF files. It was then processed based on the coordinates of the closest GRDC station and stored in CSV files for easier integration into R coding for the RF algorithm. If there are gaps in the data for the area upstream, it can result in incomplete values for the entire drainage basin. The following calculations have been performed to address missing data:

$$U_i = \sum_i V_i^*.A_i^* \tag{2.2}$$

where:

- V_i: Value of each cell in the satellite data

- V_i^*: For missing values, this sets the cell value to zero
- A_i^*: For missing values, this sets the cell area (weight) to zero
- U_i: Calculated upstream average value for each cell

An R script provided by Collot d'Escury[20] (2023) was used to identify missing values among the stations. Stations in all the datasets that did not have at least one month of data were excluded from the study. Normalized parameters are utilized as input forcing data for the PCR-GLOBWB model and also act as predictors for the Random Forest. The output state variables of PCR-GLOBWB, which were standardized to the upstream area, were employed as an RF predictor. Additionally, streamflow observations from GRDC and streamflow predictions from PCR-GLOBWB are converted into flow depth by dividing them by the catchment area. These values are then integrated as predictors for RF. Afterward, a single RF is trained using these predictors to generate a corrected value of streamflow at previously unseen locations [19].

2.2 Methods

The current study utilizes a hybrid modeling approach proposed by Magni et al. [19], with a main focus on utilizing satellite-based data to analyze the impact of remote-sensing products. Specifically, the study applies the PCR-GLOBWB hydrological model and the Random Forest method to address the research question. Streamflow predictions and state variables from PCR-GLOBWB serve as predictors for the Random Forest algorithm. Additionally, meteorological data, catchment attributes, and satellite data are used in various combinations as RF predictors. A comprehensive list is available in the Appendix 5. The next chapter will provide a detailed explanation of the models employed for hybrid modeling.

2.2.1 PCR-GLOBWB

The global hydrological model PCR-GLOBWB is a grid-based model of global terrestrial hydrology developed to assess the impact of global changes on the world's water resources. The model has five main hydrological modules: meteorological forcing, land surface, groundwater, surface water routing, and irrigation and water use, which can simulate soil moisture, snowpack, evaporation, runoff, and water storage, and it incorporates detailed representations of water flow including surface runoff, interflow, and baseflow, as well as the routing of surface water [18]. The forcing inputs of the model include precipitation, temperature, humidity, wind speed, and radiation obtained from climate models or reanalysis data such as W5E5 [30], along with land use and land cover data, water use information, and irrigation.

In this study, PCR-GLOBWB was run without calibration at 30 min resolution between 2000 and 2019 at daily timesteps. The model output was then upscaled to monthly average timesteps.

2.2.2 Random Forest

Breiman's random forest is a machine-learning algorithm that involves the creation of multiple decision trees by randomly selecting a subset of features from the original dataset [31]. It is an ensemble of trees constructed from a training data set and internally validated to predict the response based on the predictors. RF uses Gini impurity reduction for splitting, with predictors selected from a randomly chosen subset at each split, each tree is built from a bootstrap sample drawn with replacement, and predictions are aggregated through majority voting [32].

Random Forest generates multiple subsets of the original data through bootstrap sampling, where data points are randomly selected with replacement [33]. These subsets are then used to construct decision trees. The mtry parameter determines how many features are considered at each split, influencing the randomness and diversity among the trees [34]. A smaller mtry value increases randomness, while a larger value can improve accuracy but also increases the risk of overfitting due to irrelevant features. Each decision tree is built by selecting features from the subset and finding the best split at each node. The ntrees parameter controls the number of trees created in the ensemble, with more trees typically improving performance but also increasing computational costs. During tree construction, the nodesize parameter ensures that each split requires a minimum number of samples, preventing trees from becoming too deep and overfitting. However, setting nodesize too high may result in underfitting, where trees fail to capture complex relationships in the data. Once all trees are constructed, their predictions are combined through majority voting for classification tasks or averaging for regression tasks, resulting in the final prediction. This ensemble approach helps mitigate individual tree biases and produces more robust predictions. In summary, Random Forest utilizes bootstrap sampling, feature randomness, and ensemble averaging to create a collection of decision trees that collectively provide accurate and stable predictions.

In Random Forest, for each tree, a test set—disjoint from the training set—is obtained, and averaging over all these left-out data points and over all trees is known as the out-of-bag error estimate [35]. These OOB observations are utilized to estimate the model’s error. This method is beneficial for estimating test error when bagging on large datasets, where cross-validation would be computationally burdensome [36].

2.3 Modeling setup

2.3.1 Model configurations

The study involved various configurations, beginning with the work of Magni et al. We used their dataset as a benchmark for our comparative analysis to assess the impact of including satellite-based data. To study the impact of satellite-based data in a hybrid setting, we have developed several dataset configurations.

Different configurations have taken place in this study, starting with Magni et al.'s work, where we consider this dataset a benchmark for our comparative analysis to determine the effects introduced by including satellite-based data. We are conducting an investigation into the impact of satellite-based data in a hybrid setting. To achieve this, we have developed several dataset configurations.

The first configuration, pcr, utilizes the original dataset used by Magni et al. [19]. This includes PCR-GLOBWB variables, meteorological data, and catchment attributes. The second configuration, pcr_sat_add, expands upon the baseline dataset by incorporating remotely sensed variables such as satellite-based precipitation and satellite-based evaporation data. This modification allows us to analyze how additional satellite-derived information affects the modeling process. The third configuration, pcr_sat, exclusively replaces the related variables in the baseline dataset with satellite-based data. This setup enables us to isolate the impact of satellite data from other environmental parameters.

We have developed several dataset configurations to evaluate the Random Forest algorithm's performance without physically based modeling components. The fourth configuration, sat_meteo, focuses exclusively on meteorological inputs and satellite-based variables, excluding PCR-GLOBWB variables and streamflow predictions. The fifth configuration, sat_meteo_static, integrates catchment characteristics into the sat_meteo configuration to assess its performance when incorporating these specific attributes. Finally, the sixth configuration, meteo_static, includes only meteorological variables and catchment attributes. This configuration allows us to see the RF-based model performance without the contribution of satellite-based data.

Global runs and local runs are conducted for all the configurations. Local runs were conducted in Australia, Brazil, Canada, Russia, South Africa, and the United States. South Africa, Brazil, and Canada were chosen due to their distinct climate characteristics in order to assess the model's performance across varied climatic conditions. These countries have similar data availability as shown in Figure 2.1. Additionally, Australia, Russia, and the United States were selected based on differences in data availability to analyze how the model performs with varying levels of data accessibility.

After the model was trained using both global and local data, its performance in South Africa, Brazil, Canada, Australia, Russia, and the United States was compared to determine whether global or local training data yields better results in these countries.

Table 2.1: Predictors used in different model configurations. * indicates the exclusion of precipitation and total evaporation.

Configuration	PCR-GLOBWB discharge and state variables	Meteorological variables	Catchment Attributes	Satellite Data
pcr	X	X	X	
pcr_sat_add	X	X	X	X
pcr_sat	X	X	X	X*
sat_meteo		X		X
sat_meteo_static		X	X	X
meteo_static		X	X	

2.4 Model training and evaluation

Random Forest is a machine learning algorithm that has been employed in R using the Ranger package. The Ranger package is known for effectively managing extensive datasets and impressive performance [37]. Optimizing hyperparameters such as the number of trees, mtry, and nodesize can significantly improve the performance of the algorithm. This process tailors the model to meet the specific requirements and objectives of the dataset. Properly selecting and adapting hyperparameters can increase the model's precision, resilience, and overall learning abilities.

This study adjusted mtry while keeping the number of trees at 200 and the nodesize at 5. The Out-Of-Bag Root Mean Squared Error metric was used to determine the optimal value for mtry in the model. In this approach, the model's performance was evaluated with different values of mtry, and the value that yielded the lowest error rate was selected.

A location-based split sampling was utilized to train and validate the Random Forest model. This process involves dividing the data into five subsamples based on location. Each subsample is then further divided into training and testing stations. Within each subsample, 70% of the data was used to train the model, and the remaining 30% was used to test its accuracy. This rigorous method ensures that the model is trained on diverse data, representing the entire dataset and not just a specific subset. The entire training data set has been merged and organized into a single table, which will be used to train the RF (Random Forest) model. The exact process has also been applied to the testing data. Once the model was trained, it generated new predictions for every station in the testing dataset.

The model's performance was evaluated using the Kling-Gupta Efficiency metric, which combines various parameters for the calculation. The Kling–Gupta Efficiency (KGE) is used in hydrological modeling to evaluate the accuracy of simulated hydrographs compared to observed data. The KGE combines three components, correlation, bias, and variability, into a single efficiency measure [38].

The formula for KGE is:

$$KGE = 1 - \sqrt{(r-1)^2 + (a-1)^2 + (b-1)^2} \tag{2.3}$$

where:

- r is the correlation coefficient between observed and simulated values,
- a is the ratio of the standard deviation of the simulated values to the standard deviation of the observed values,
- b is the ratio of the simulated values' mean to the observed values' mean.

KGE ranges from -∞ to 1, with 1 indicating a perfect match between simulated and observed data [39]. A higher KGE value signifies better model performance, taking into account correlation, bias, and variability simultaneously.

3. Results

This chapter is segmented into two parts: global runs and local performance. The findings from these sections will be elucidated in distinct subsections.

3.1 Global runs

3.1.1 Hyperparameter tuning

Figure 3.1 displays the tuning methodology implemented across all configurations and for every one of the five subsamples. The number of trees (ntree) was maintained at a total of 200. The range of adjusted values of mtry fluctuated based on the count of predictors present. The objective behind tuning was to ascertain an ideal mtry parameter value that reduces out-of-bag root mean square error (00B RMSE), consequently improving model efficacy. Due to computational constraints, only the mtry parameter has been tuned as it is relatively less computationally demanding. Both meteo_static and sat_meteo_static show similar graphs with a significant drop in error in the first few values, followed by minimal fluctuations after the minimum 00B RMSE value reached, and close values between all of the subsamples. Tuning graphs of pcr, pcr_sat and pcr_sat_add show greater fluctuations after the minimum OOB RMSE value. The difference between subsamples are also greater in these configurations. The configurations meteo_static and sat_meteo_static exhibit similar graphs with fewer fluctuations beyond the minimum 00B RMSE value and closely matched values across all subsamples. The pcr, pcr_sat, and pcr_sat_add tuning graphs reveal more significant fluctuations after reaching the minimum 00B RMSE value. Additionally, variations between subsamples are notably larger in these configurations. This is coupled with a marginal elevation in 00B RMSE across these setups, which might indicate an overfitting phenomenon within the training dataset. Since sat_meteo was only calibrated for a few variables, comparing it with other charts would be uninformative. The optimal values of mtry for every configuration are summarized in Table 3.1.

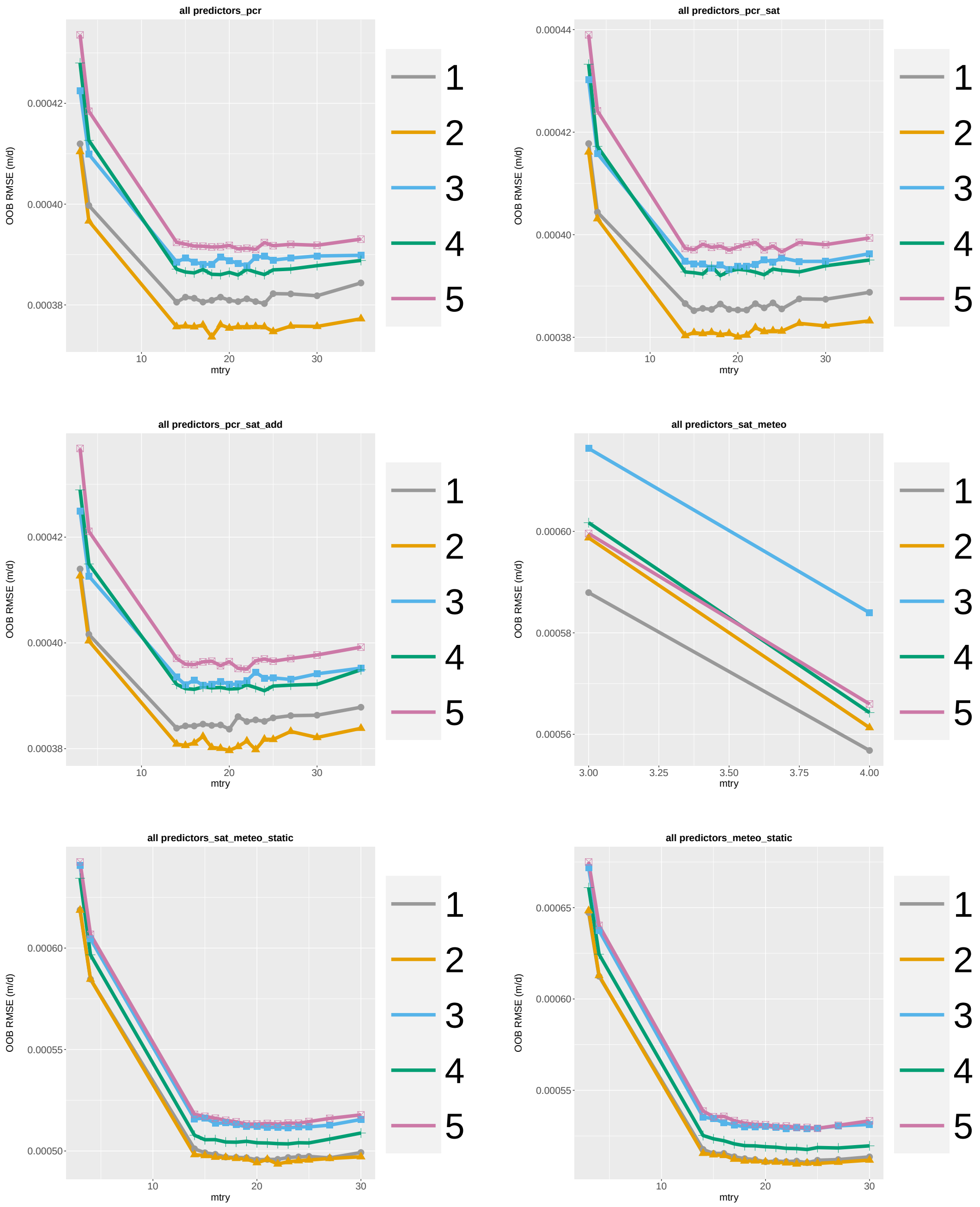


Figure 3.1: RF tuning of the mtry hyperparameter, with each panel displaying all subsamples and a single configuration. The dots indicate the tuned mtry values, while the lines denote the OOB RMSE score. A fixed ntree of 200 and a node size of 5 was used.

	pcr	pcrSatAdd	pcrSat	satMeteo	satMeteoStatic	meteoStatic
S1	24	20	15	4	20	24
S2	18	20	20	4	22	23
S3	22	17	19	4	23	24
S4	21	24	18	4	23	24
S5	23	22	25	4	19	25

Table 3.1: mtry values for each subsample of each configuration.

3.1.2 Variable importance

Figure D.12 presents the mean decrease in impurity values for the top twenty variables across all five global RF. The variable importance analysis was conducted to determine the relative importance of each predictor in the model. pcr, pcrSatAdd, and pcrSat demonstrate comparable trends in terms of variable significance. It is evident that satellite-based evaporation ranks third in importance among the configurations incorporating satellite data. Static variables make up nearly half of the top 20 variables. In configurations excluding PCR-GLOBWB variables, meteoStatic and satMeteoStatic yield very similar outcomes, with the aridity index as the top variable followed by all the meteorological input. However, with the inclusion of satellite data, evaporation emerges as the second most crucial factor after aridityIdx. When using both meteorological and satellite data (i.e., SatMeteo), it is apparent that satellite-based evaporation holds a prominent position in variable importance by ranking at the top. Satellite-based precipitation does not exhibit an equivalent level of significance as evaporation across both included configurations.

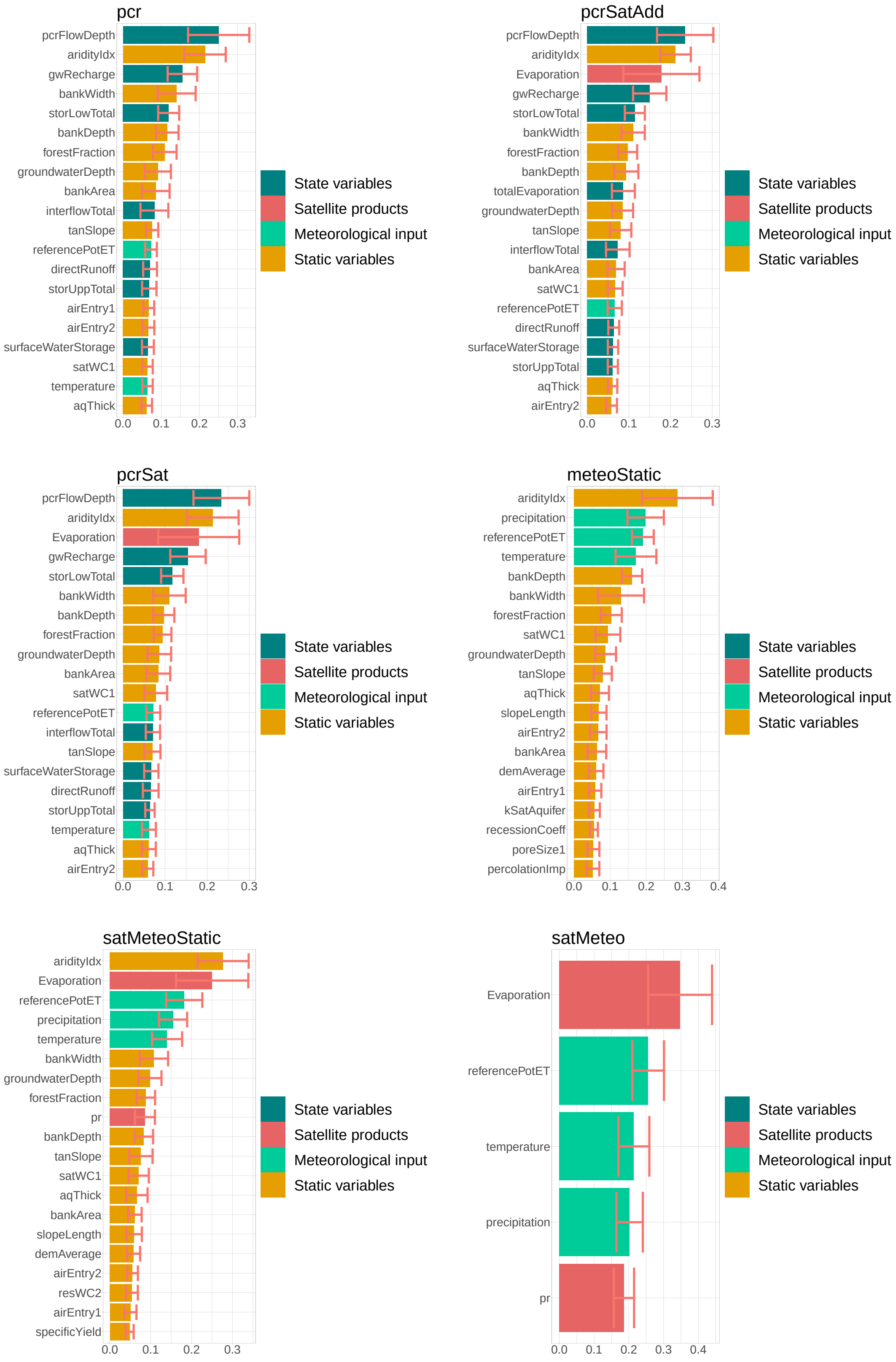


Figure 3.2: Square rooted mean decrease in impurity values of the top twenty variables, averaged over five training subsamples for all the global Random Forest (RF) configurations. Each type of variable is represented by a different color.

3.1.3 Performance

Figure 3.3 depicts the KGE values of all configurations in boxplots, with each column representing five subsamples and their cumulative distribution. The model exhibits similar performance across all subsamples for all configurations. This consistency indicates the potential for generalization beyond the specific samples it was trained on. The pcr, pcrSat, pcrSatAdd, MeteoStatic, and SatMeteoStatic configurations consistently yield successful results with KGE values closer to 1 and smaller boxplots. In contrast, the satMeteo and uncalibrated configurations exhibit poorer performance, with satMeteo being particularly inferior between the two. The introduction of static variables to the satMeteo configuration significantly improves the model's performance, as indicated by the increase in KGE value from satMeteo to satMeteoStatic.

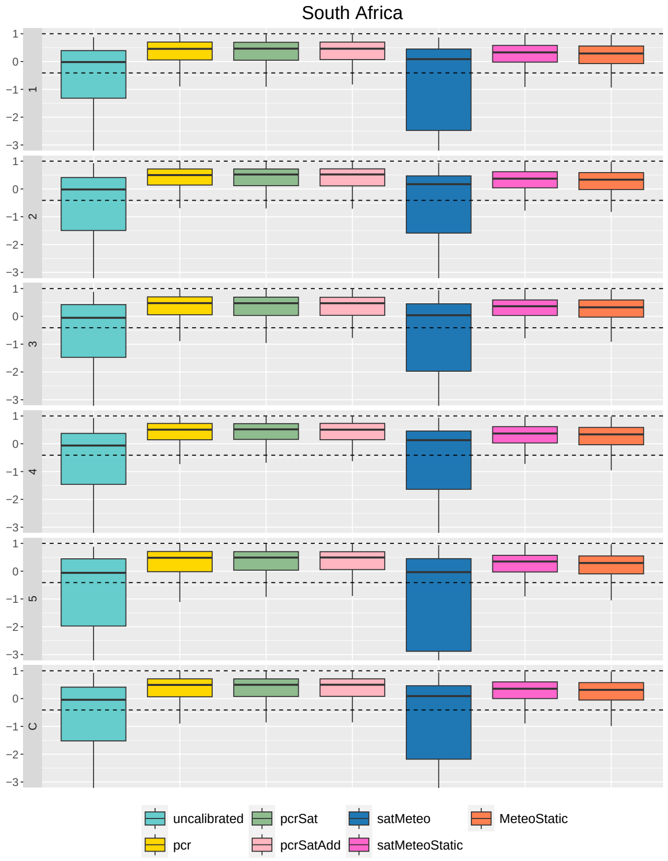


Figure 3.3: Boxplots of KGE for all five subsamples and their accumulation as rows for six different configurations, including uncalibrated PCR-GLOBWB discharge simulations. The dashed lines denote ideal (1) and 'good' (-0.41) KGE values.

Figure 3.4 shows the cumulative KGE values of all the configurations averaged over all subsamples.

pcr, pcr sat, pcr sat add,sat meteo static and meteo static have all delivered similar performances. All of their results are better than the uncalibrated PCR-GLOBWB and sat_meteo configuration. Among the well-performing configurations, sat_meteo_static and meteo_static demonstrate marginally better results until KGE values reach 0.

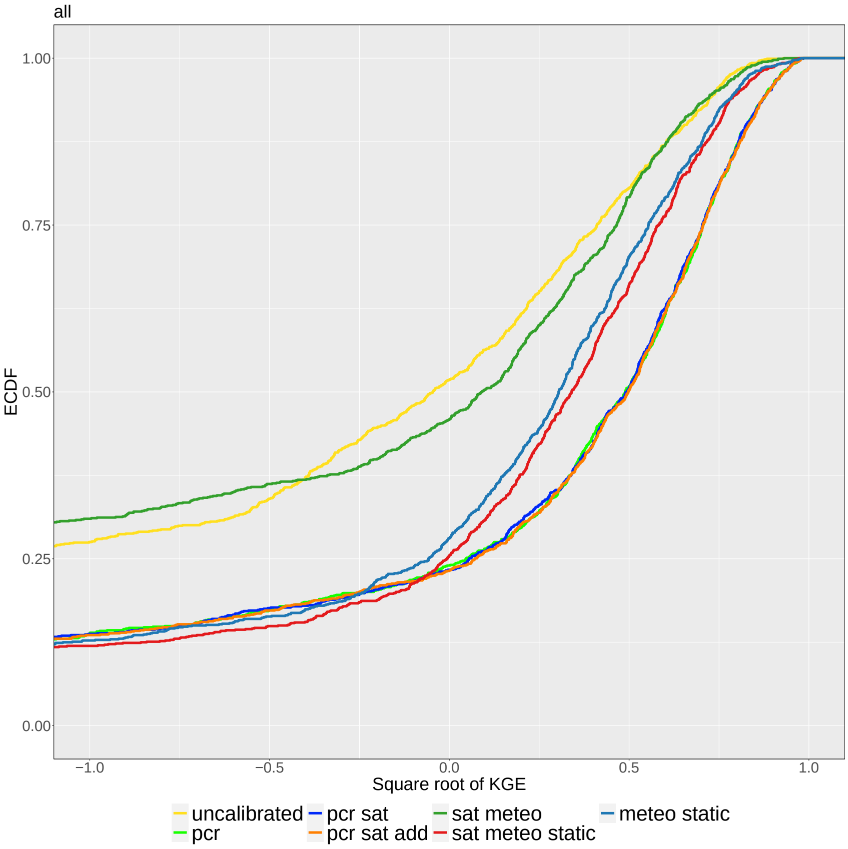


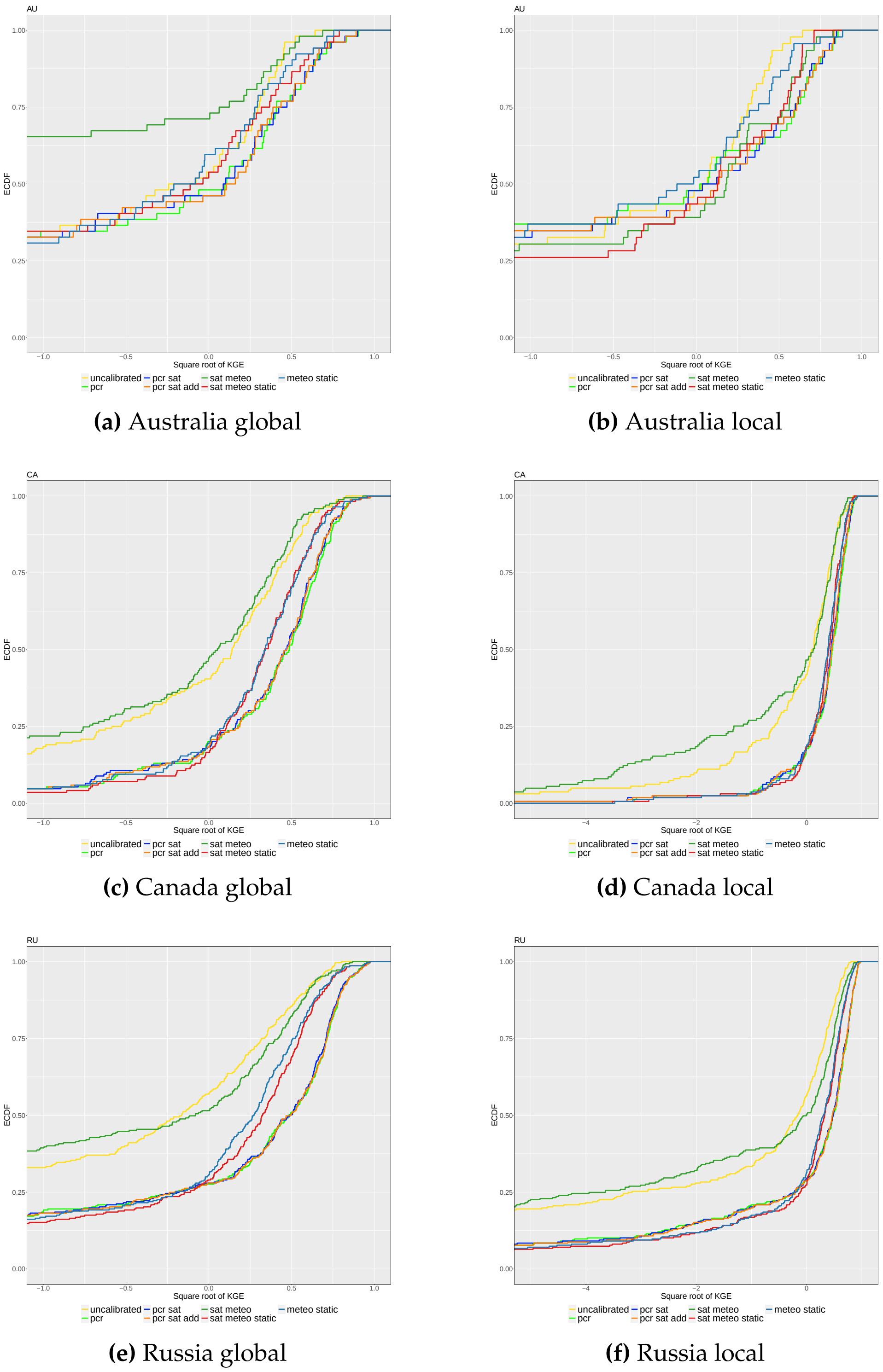
Figure 3.4: Cumulative distribution functions of KGE for the six configurations and the uncalibrated PCR-GLOBWB, with KGE results averaged across five subsamples. Only higher KGE scores are depicted, with the x-axis constrained to -5.

3.2 Local runs

Local runs have been conducted in Australia, Brazil, Canada, Russia, the United States, and South Africa to assess the model's performance in countries with varying climatic conditions and different levels of data availability. The performance of the model was evaluated after global and local training in these countries.

3.2.1 Performance after global training vs. local training

The cumulative distributions of KGE values from 5 subsamples of all configurations for the countries - Brazil, Russia, and the US - indicate that global training of the model demonstrates similar performance to locally trained models in these countries. Specifically for Russia, configurations consisting of PCR-GLOBWB variables, including uncalibrated configuration, exhibit better performance in globally trained models compared to other configurations, which show similar performance for both runs. In Canada's case, uncalibrated and sat_meteo configurations performed slightly better in global training while remaining configurations showed similar performances for both runs. Australia shows comparable performances between uncalibrated, pcr_sat, and pcr_sat_add configurations for both global and local training. However, sat_meteo and sat_meteo_static displayed worse performance in the global run while meteo_static configuration showed better performance in local run. Graphs of the cumulative distributions of KGE values for all countries, along with KGE boxplots, are available in the Appendix 5.



(a) Australia global

(b) Australia local

(c) Canada global

(d) Canada local

(e) Russia global

(f) Russia local

Figure 3.5: Cumulative distribution functions of globally trained KGE for the six configurations and the uncalibrated PCR-GLOBWB, with KGE results averaged across five subsamples for Australia, Canada, and Russia. Only higher KGE scores are depicted, with the x-axis constrained to -5.

3.2.2 Different data availability

The figure 3.6 presented in this analysis shows the KGE boxplot of Australia, Russia, and the United States. These countries have varying data availability. In Australia, sat_meteo_static shows the best performance in all five subsamples. This indicates that the satellite-based dataset significantly contributes to the accuracy of the model. This improvement is particularly noticeable when comparing sat_meteo_static to meteo_static, with satellite data contributing to both higher accuracy and greater consistency in a RF-based model, as evidenced by a narrower boxplot. For the US and Russia, configurations including pcr variables marginally outperform other setups. Importantly, both countries exhibit marked improvements in model performance when catchment attributes are incorporated into the setups, underscoring the value of catchment characteristics. Russia's performance surpasses that of Australia but falls short of the US. As data availability increases, there is a clear trend towards improved model performance and consistency.

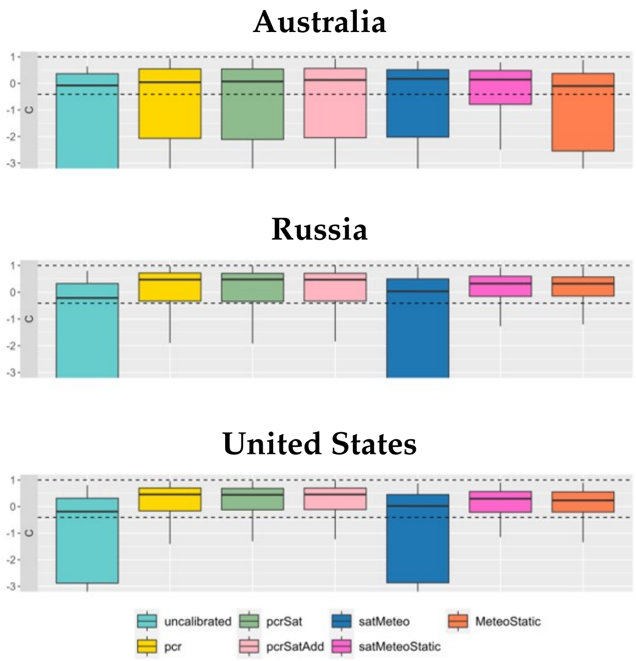
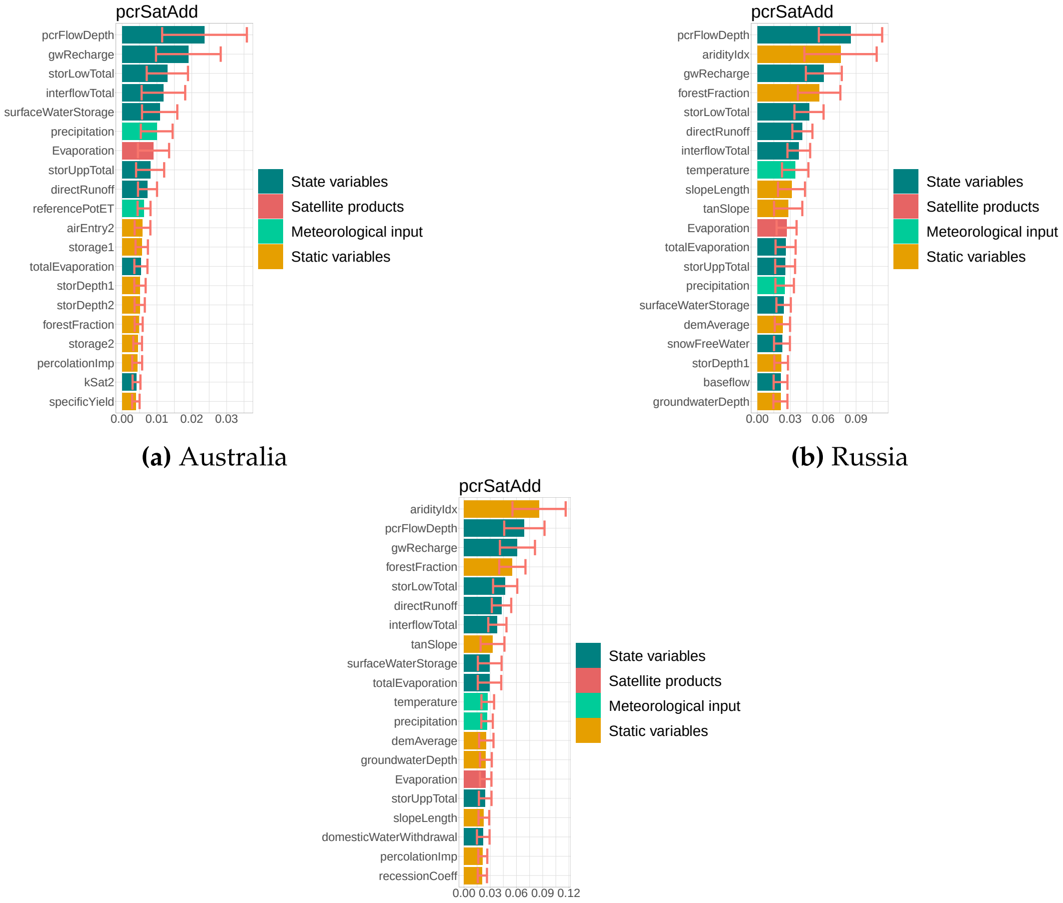


Figure 3.6: Boxplots of KGE values accumulated over five subsamples for six different configurations, including uncalibrated PCR-GLOBWB discharge simulations for Australia, Russia, and the US. The dashed lines denote ideal (1) and 'good' (-0.41) KGE values.

The variable importance graphics display the relative significance of input variables in influencing the performance of the hydrological model. In Australia, the top 5 prominent variables consist of hydrological state variables followed by precipitation and satellite-based evaporation. Similarly, in Russia, these five variables are also quite important; however, aridity index and forest fraction also take second and fourth places respectively with temperature from meteorological variables coming after. In the United States, the most significant variable becomes aridity index but the top 8 variables are very similar to those in Russia's graph. Both temperature and precipitation come after these key factors. As data availability increases from Australia to Russia, and then from Russia to the US, satellite-based evaporation becomes less significant while static variables become more critical.



(a) Australia

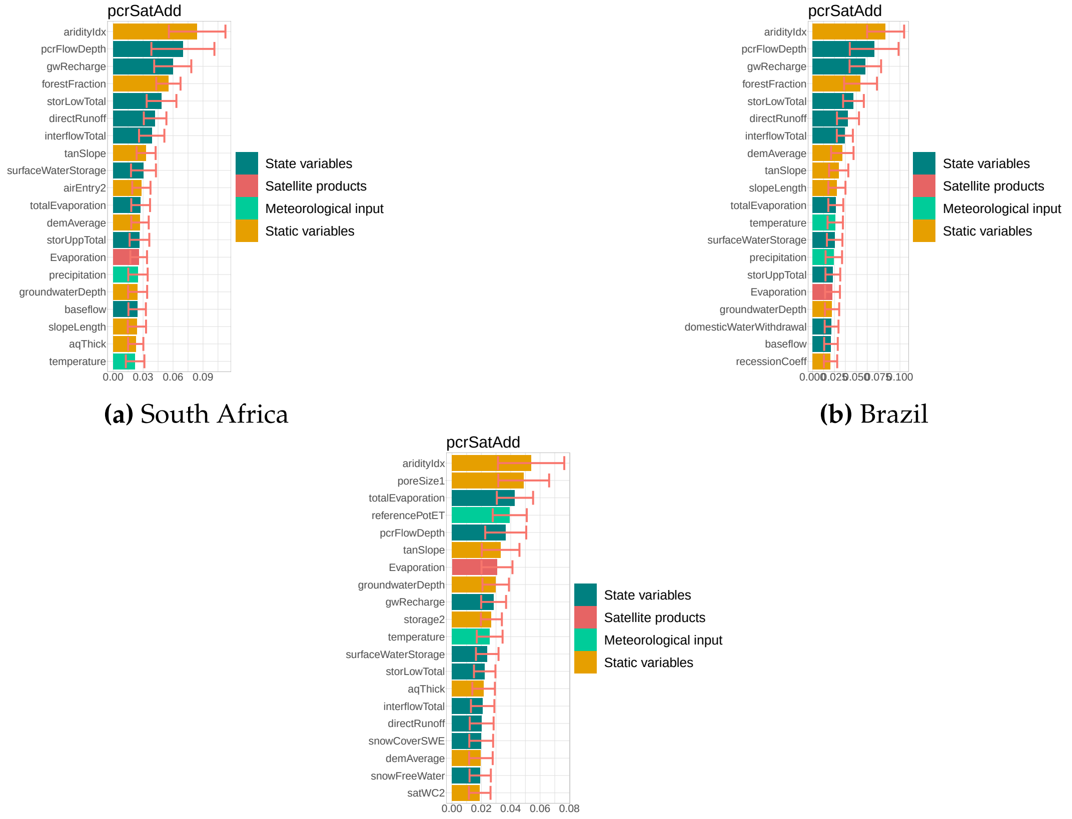
(b) Russia

(c) US

Figure 3.7: Square rooted mean decrease in impurity values of the top twenty variables, averaged over five training subsamples for pcrSatAdd Forest (RF) configuration for Australia, Russia and the US. Each type of variable is represented by a different color.

3.2.3 Different climatic regions

Aridity index is the most important variable for all three types of climatic countries. Hydrological state variables consist almost half of the top 20 variables for all of three countries, however their variations change for each of the countries. South Africa and Brazil show similar trends in the top 8 variables. While satellite-based evaporation gets place in the top 20 important variables with different significance levels, satellite-precipitation didn't get into the list in any of them. Satellite-based evaporation is highly important for Canada's case.



(a) South Africa

(b) Brazil

(c) Canada

Figure 3.8: Square rooted mean decrease in impurity values of the top twenty variables, averaged over five training subsamples for pcrSatAdd Random Forest (RF) configuration for South Africa, Brazil, and Canada. Each type of variable is represented by a different color.

Figure 3.9 illustrates the cumulative KGE values for three countries: South Africa, Brazil, and Canada. Overall, Canada demonstrated the best performance among these countries. The models pcr_sat_add, pcr, pcr_sat, sat_-meteo_static, and meteo_static all exhibited strong performances for all of the countries. Additionally, both sat_meteo_static and meteo_static showed slightly better performance in line with global runs. It is worth noting that Canada outperformed South Africa and Brazil in uncalibrated and sat_meteo configurations.

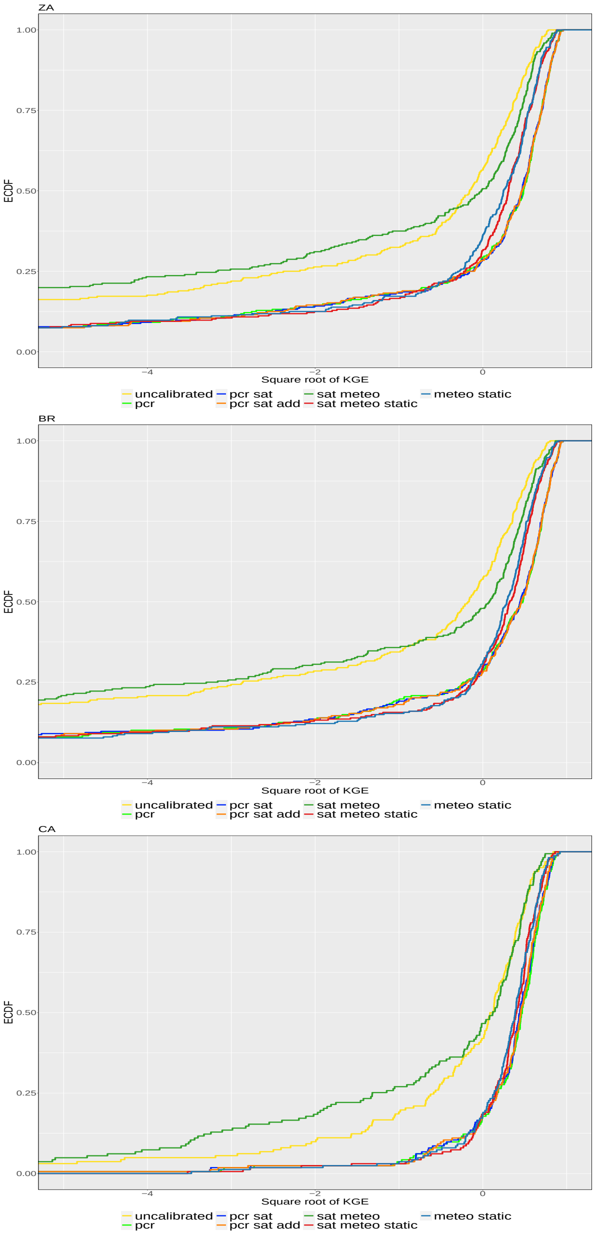


Figure 3.9: Cumulative distribution functions of KGE for the six configurations and the uncalibrated PCR-GLOBWB, with KGE results averaged across five subsamples for South Africa, Brazil, and Canada. Only higher KGE scores are depicted, with the x-axis constrained to -5.

4. Discussion

4.1

This study investigated the impact of integrating satellite-based precipitation and evaporation data into a hybrid hydrological model to predict streamflow across various configurations, climates, and regions, emphasizing the crucial role of satellite data in improving accuracy and reliability. Incorporating satellite-based precipitation and evaporation data did not result in a significant enhancement of global streamflow prediction accuracy. Nonetheless, in regions where data is sparse, there was a slight improvement in the model's efficiency. Although certain stations experienced an increase in KGE values, the overall performance of the model was not significantly improved by the utilization of satellite-based data. This suggests that while satellite data can be informative, its influence on model performance depends on the context and may only lead to enhanced model performance in some instances. The study shows that removing the PCR-GLOBWB inputs does not significantly affect model performance, allowing for more flexible model configurations when PCR-GLOBWB data may be limited or unavailable. The model's performance, when utilizing a combination of satellite data, meteorological information, and catchment characteristics, was notably comparable to scenarios where only PCR-GLOBWB inputs were used, and it significantly outperformed the uncalibrated PCR-GLOBWB configurations. This outcome signifies a potentially more practical and efficient approach to hydrological modeling. However, it is crucial to recognize this methodology's limitations and contextual nuances. Firstly, the reliance on remote sensing data introduces constraints, particularly concerning future projections. Since remote sensing data captures current and historical conditions, its applicability in forecasting future scenarios is limited. Moreover, while a Random Forest model, trained on observational data, may demonstrate high performance in simulating known conditions, it may struggle to accurately predict outcomes under unobserved scenarios. Unlike process-based models, RF models do not inherently inform on the underlying mechanisms of streamflow. Instead, they offer a statistical representation based on historical data, which may not capture the full complexity of hydrological processes or the impacts of novel conditions on these processes. An extensively calibrated PCR-GLOBWB model could potentially exhibit improved performance over an RF model, highlighting the importance of model calibration in achieving optimal results. This observation challenges the conventional reliance on comprehensive global ground-based hydrological datasets but also underscores the importance of understanding the context-specific applicability of alternative data integration approaches.

Model performance, when trained on local or global data, significantly depends on the region's specific characteristics and data availability. Globally trained models use broader PCR-GLOBWB variables and satellite data datasets to capture wide-ranging patterns effectively. However, the locally trained model that excluded PCR-GLOBWB variables demonstrated superior performance in Australia, a data-scarce region, by heavily relying on satellite data to capture region characteristics and climatic conditions that improve predictions. An earlier study found that the hybrid model can perform similarly to physically based models at a global level but achieved better local adaptivity [13]. In our case, the local runs using the hybrid modeling approach performed similarly to global runs, and in some regions, they even performed worse. For Russia, global training of the hybrid model demonstrated better performance. This observation highlights a trade-off between globally and locally trained models: while globally trained ones offer generalization capabilities across diverse environments, locally trained ones provide enhanced precision for specific contexts.

The hybrid model shows better performance as the data availability increases. The study indicates a significant enhancement in the model's performance in Australia when focusing on satellite-based data, meteorological data, and catchment attributes. In a previous study [20], local runs for Australia were also conducted, which included several other satellite products but did not incorporate PCR-GLOBWB variables and catchment attributes; the results did not show any significant improvement in the model's performance in Australia. This underscores the significance of choosing specific satellite data and incorporating the characteristics of the study area to improve hydrological predictions. The superior performance in Australia suggests satellite-based inputs can offer a more precise and dependable prediction model for regions with certain characteristics.

The model performance varies across Canada, Brazil, and South Africa, underscoring the influence of regional characteristics on hydrological modeling. Factors such as climatic diversity, data availability, hydrological and geographical features contribute to this variability. In addition to varying model performance, satellite-based data showed great significance in the variable importance list but did not affect the model's performance. Integrating satellite data can be particularly beneficial for helping to capture a wide range of hydrological processes. However, the overall impact of satellite data may be limited if the model is already well-calibrated with high-quality ground-based data. Several studies have evaluated the hydrological applicability of different satellite-based data and have generally found that these data are less effective as inputs compared to ground observations [40]. This highlights the need for model structures and calibration techniques that effectively utilize the unique strengths of satellite observations.

4.2 Scientific and societal impacts

Using satellite-based precipitation and evaporation data in hydrological models represents a big step forward for the field of hydrology. Valuable information that satellite data provides as an alternative data source bolsters our capacity to understand and respond to existing water availability and hydrological trends. This new approach takes us one step closer to solving problems regarding model performance espescially in areas with limited ground-based observations. The success of using satellite data in regions where monitoring is scarce provides a strong tool for model performance but also sets a new standard for future research on using remote sensing technologies to model the environment.

Information that satellite data offers affects the practices in agriculture as improved model performance can inform better irrigation practices, contributing to water conservation and ensuring crop viability. In urban planning, accurate models can guide infrastructure development to manage stormwater effectively, reducing the risk of flood damage and improving water quality. Moreover, while satellite data may not predict future events, its detailed insights into past and present hydrological conditions are invaluable for developing more robust risk assessment tools. By understanding the variability and trends in water resources over time, communities can better anticipate potential water shortages or surplus conditions, aligning water use policies and conservation efforts more closely with the realities of their hydrological environment. This approach enhances resilience to climate variability, supports the sustainable management of water resources, and contributes to the overall well-being and safety of populations.

The methodology employed in this research, which harmonizes traditional hydrological models with machine learning techniques and satellite data, has broader implications beyond hydrology. Similar approaches could revolutionize data integration and modeling in related fields such as climatology, environmental science, and agriculture. For instance, climate models could benefit from enhanced predictive accuracy by incorporating satellite observations directly into model calibration processes, leading to improved climate projections and mitigation strategies. In agriculture, predictive models that accurately forecast water availability can inform irrigation planning, crop selection, and drought management practices, contributing to more sustainable agricultural practices. These potential applications show how hybrid modeling methods can be applied in various scientific fields with great flexibility and transformative power.

4.3 Limitations and recommendations

Many other satellite data sources can improve streamflow predictions in our hybrid modeling setup, where we combine PCR-GLOBWB with the RF method. This could involve using better-quality variables or adding more data to supplement what is already available. Research has shown that integrating remote sensing LAI into the PCR-GLOBWB model improves evapotranspiration and discharge estimates, leading to better overall performance [41]. Including LAI data provides valuable insights into vegetative cover in a watershed area, resulting in a more comprehensive representation of hydrological processes and improved streamflow estimations. Furthermore, a previous study by Collot d'Escury [20] incorporated Liquid Water Equivalent, Snow Cover Fraction, and Soil Moisture data into the same hybrid modeling setup as our study and demonstrated promising results. Therefore, integrating these datasets with our study has the potential to significantly enhance the model's performance.

The model's performance varies across regions with different climates and geography, showing that the model may adapt poorly to certain hydrological processes and conditions. These regions have distinct climatic conditions, which indicate the need for tailored models to address their specific challenges effectively. The current hybrid approach may only partially meet these unique requirements. Adjusting the calibration process for the PCR-GLOBWB model to match each region's specific local conditions—such as modifying parameters like soil moisture capacity, vegetation characteristics, and infiltration rates—could help resolve this issue.

Additionally, the selection of countries allows for the investigation of the model's performance across various climatic regions. However, these countries also exhibit diverse climatic and geographical variations within themselves. Focusing on specific regions with similar climatic conditions instead of entire countries would lead to more accurate results regarding the model's performance in this context. Clustering spatial data can identify regions with similar hydrological characteristics, land cover, and climate variables.

The research has identified a significant improvement in model performance in Australia when using satellite data, particularly for models trained locally. Adopting a transfer learning approach could be advantageous in regions where specific satellite data holds less influence and broader datasets are more representative. Transfer learning leverages information from data-rich regions to improve predictions in areas with limited available data[42]. This approach can enhance the model's performance by addressing the specific data needs of some regions with less data availability.

5. Conclusion

This study explores the effectiveness of integrating satellite-based precipitation and evaporation data into a hybrid hydrological model for streamflow prediction. The findings highlight the varying impact of satellite data on model accuracy in different global regions, suggesting contextual benefits rather than consistent improvements. The analysis reveals that in data-scarce areas like Australia, satellite data enhances model performance by capturing unique regional characteristics, emphasizing the importance of tailored model training strategies for both broad applicability and localized accuracy. Furthermore, the study suggests a potential trade-off between models trained on global datasets and those with local calibration - with superior performance shown in specific contexts by locally calibrated models. Despite varied model performance across regions, integrating satellite data is emphasized as crucial for improving outcomes in areas lacking ground-based observations. Additionally, there's a need to further explore PCR-GLOBWB model biases and limitations while considering alternative satellite data sources and transfer learning approaches to enhance streamflow predictions in diverse environments.

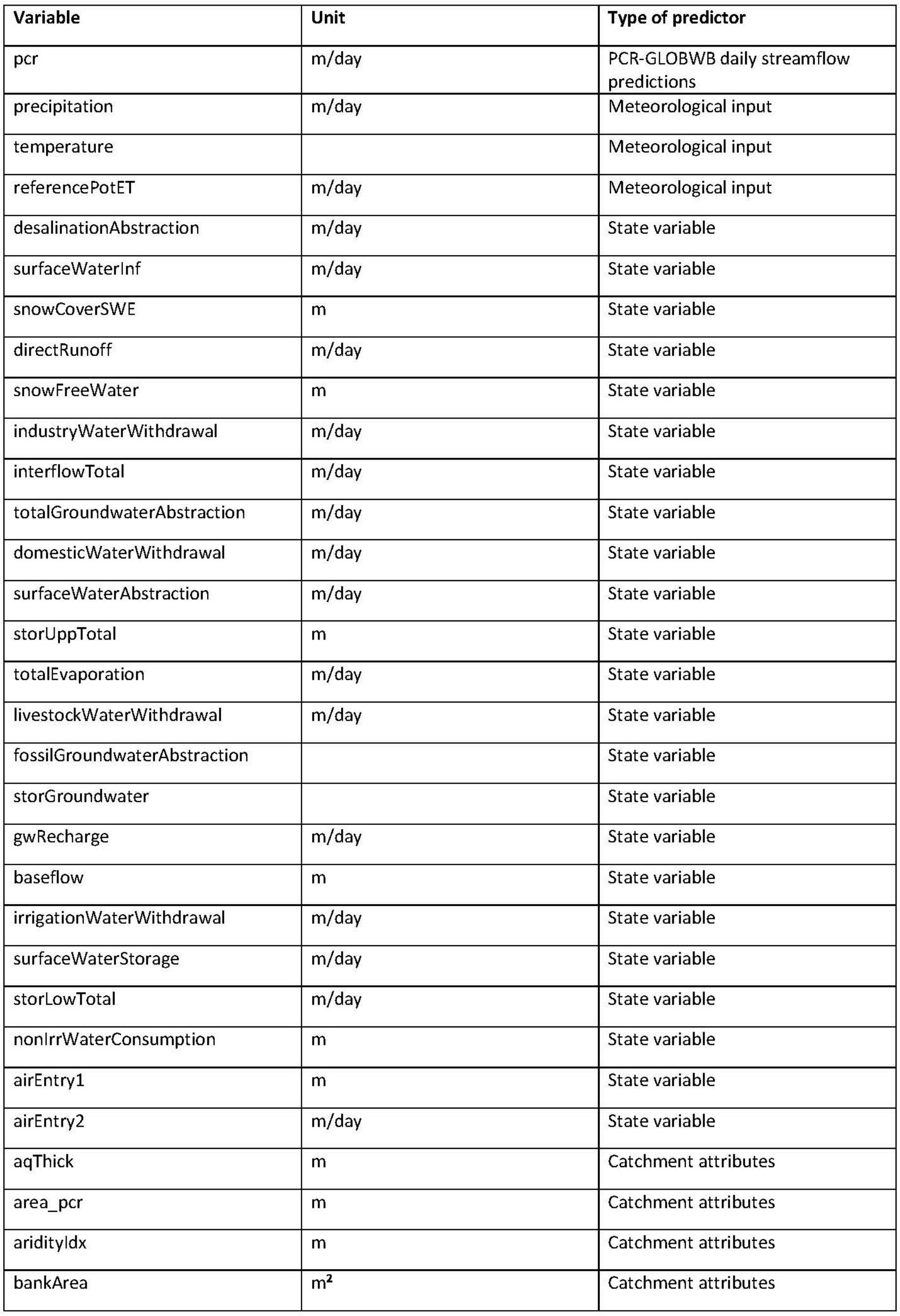
This study confirms the efficacy of incorporating satellite data into hydrological models, particularly emphasizing its role in regions lacking ground observations. This approach not only aids in capturing unique regional hydrological characteristics but also in optimizing model efficiency and applicability. The implications for hydrology and related fields are profound, suggesting a paradigm shift towards more adaptive and region-specific model calibration strategies. Such advancements promise enhanced water resource management, informed by more accurate and efficient hydrological modeling techniques, thereby contributing valuable insights into the sustainability and resilience of water systems under changing global conditions.

A. Appendix

The code associated with this study is accessible through the following link: https://github.com/sbusraisik/PCR-GLOBWB-RF-satellitedata

B. Appendix

Table A.1: PCR-GLOBWB variables full list.

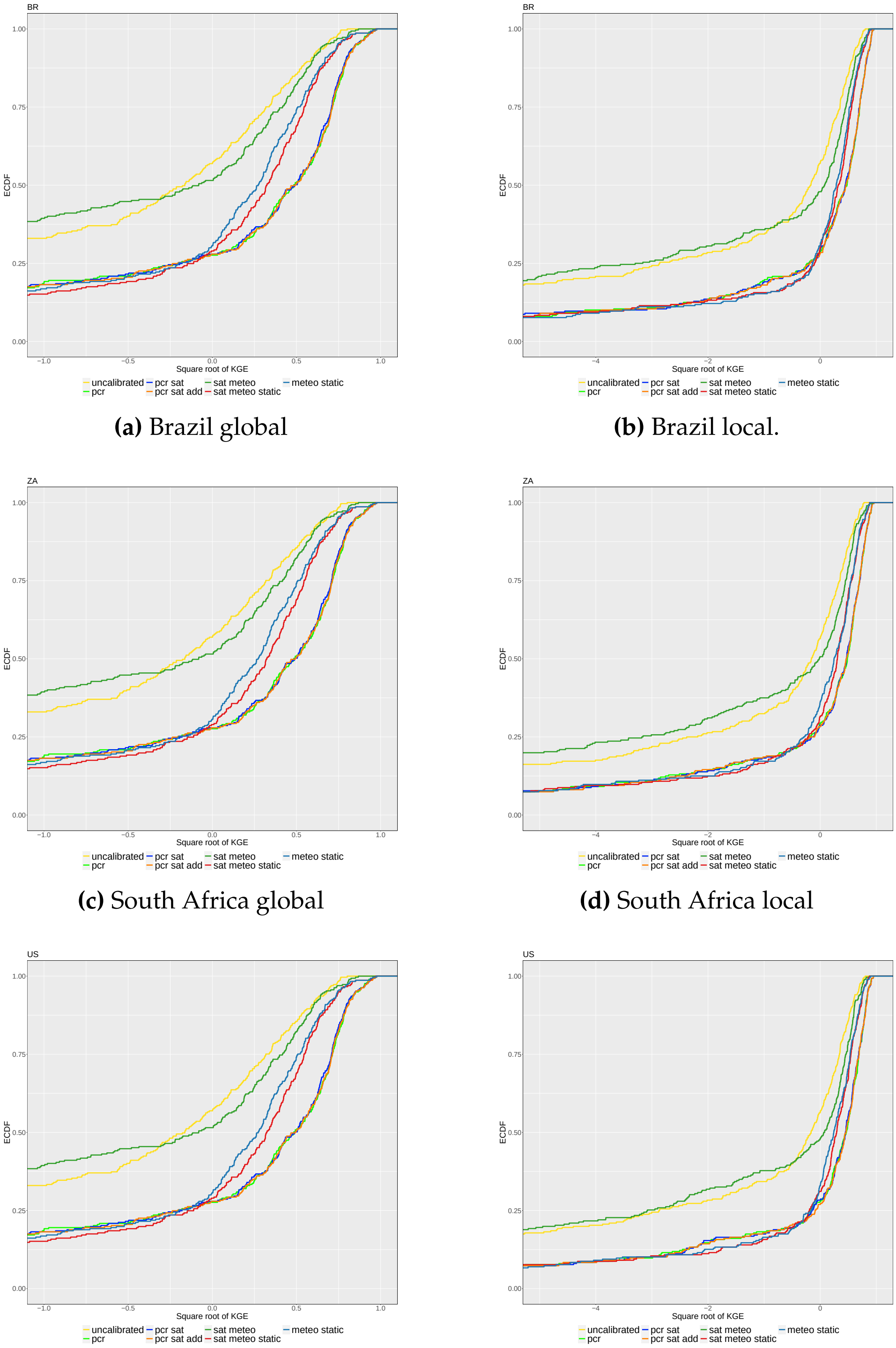


Variable	Unit	Type of predictor
pcr	m/day	PCR-GLOBWB daily streamflow predictions
precipitation	m/day	Meteorological input
temperature		Meteorological input
referencePotET	m/day	Meteorological input
desalinationAbstraction	m/day	State variable
surfaceWaterInf	m/day	State variable
snowCoverSWE	m	State variable
directRunoff	m/day	State variable
snowFreeWater	m	State variable
industryWaterWithdrawal	m/day	State variable
interflowTotal	m/day	State variable
totalGroundwaterAbstraction	m/day	State variable
domesticWaterWithdrawal	m/day	State variable
surfaceWaterAbstraction	m/day	State variable
storUppTotal	m	State variable
totalEvaporation	m/day	State variable
livestockWaterWithdrawal	m/day	State variable
fossilGroundwaterAbstraction		State variable
storGroundwater		State variable
gwRecharge	m/day	State variable
baseflow	m	State variable
irrigationWaterWithdrawal	m/day	State variable
surfaceWaterStorage	m/day	State variable
storLowTotal	m/day	State variable
nonIrrWaterConsumption	m	State variable
airEntry1	m	State variable
airEntry2	m/day	State variable
aqThick	m	Catchment attributes
area_pcr	m	Catchment attributes
aridityIdx	m	Catchment attributes
bankArea	m²	Catchment attributes

bankDepth	-	Catchment attributes
bankWidth	m^2	Catchment attributes
demAverage	m	Catchment attributes
forestFraction	m	Catchment attributes
groundwaterDepth	m	Catchment attributes
KSat1	-	Catchment attributes
KSat2	m	Catchment attributes
kSatAquifer	m/day	Catchment attributes
percolationImp	m/day	Catchment attributes
poreSize1	m/day	Catchment attributes
poreSize2	-	Catchment attributes
recessionCoeff	-	Catchment attributes
resWC1	-	Catchment attributes
resWC2	day^{-1}	Catchment attributes
satWC1	m^3/m^3	Catchment attributes
satWC2	m^3/m^3	Catchment attributes
slopeLength	m^3/m^3	Catchment attributes
specificYield	m^3/m^3	Catchment attributes
storage1	m	Catchment attributes
storage2	m^3/m^3	Catchment attributes
storDepth1	m	Catchment attributes
storDepth2	m	Catchment attributes
tanSlope	m	Catchment attributes

C. Appendix

C.1 Cumulative distribution functions



(a) Brazil global

(b) Brazil local.

(c) South Africa global

(d) South Africa local

(e) US global

(f) US local

Figure C.1: Cumulative distribution functions of globally trained KGE for the six configurations and the uncalibrated PCR-GLOBWB, with KGE results averaged across five subsamples for Brazil, South Africa, and the US. Only higher KGE scores are depicted, with the x-axis constrained to -5.

D. Appendix

D.1 KGE Boxplots of Local Runs

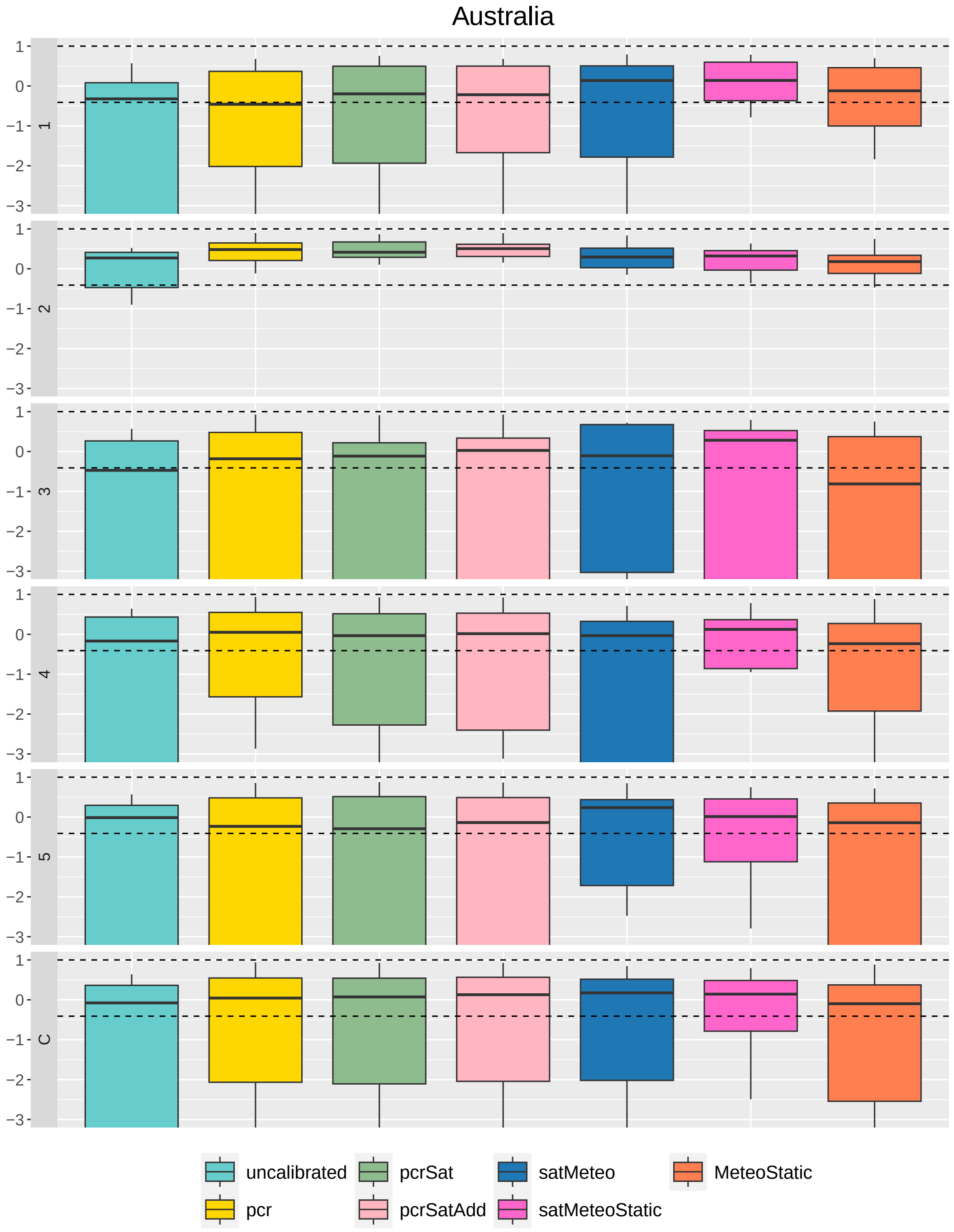


Figure D.1: Boxplots of KGE for all five subsamples and their accumulation for six different configurations in Australia, including uncalibrated PCR-GLOBWB discharge simulations. The dashed lines denote ideal (1) and 'good' (-0.41) KGE values.

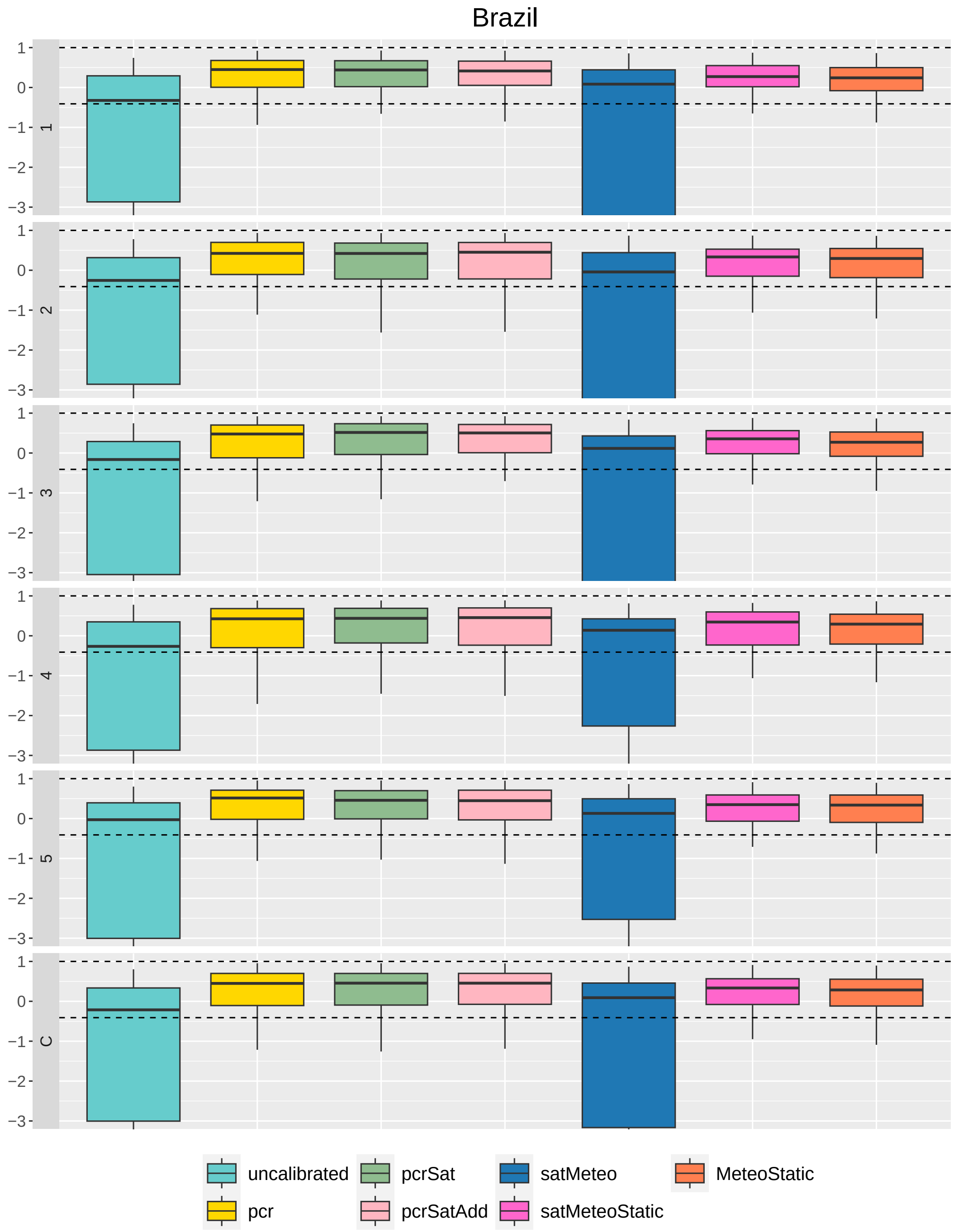


Figure D.2: Boxplots of KGE for all five subsamples and their accumulation for six different configurations in Brazil, including uncalibrated PCR-GLOBWB discharge simulations. The dashed lines denote ideal (1) and 'good' (-0.41) KGE values.

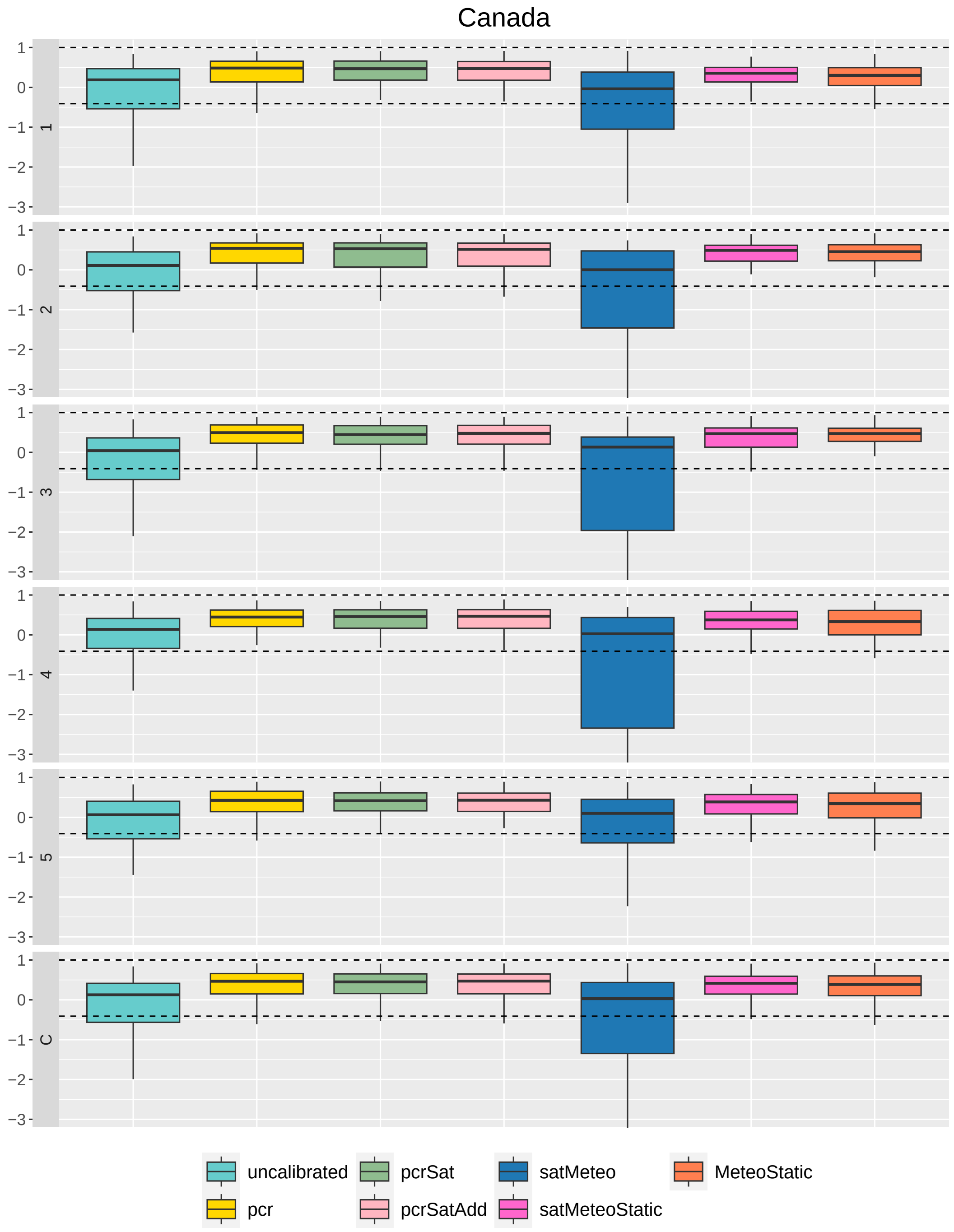


Figure D.3: Boxplots of KGE for all five subsamples and their accumulation for six different configurations in Canada, including uncalibrated PCR-GLOBWB discharge simulations. The dashed lines denote ideal (1) and ’good’ (-0.41) KGE values.

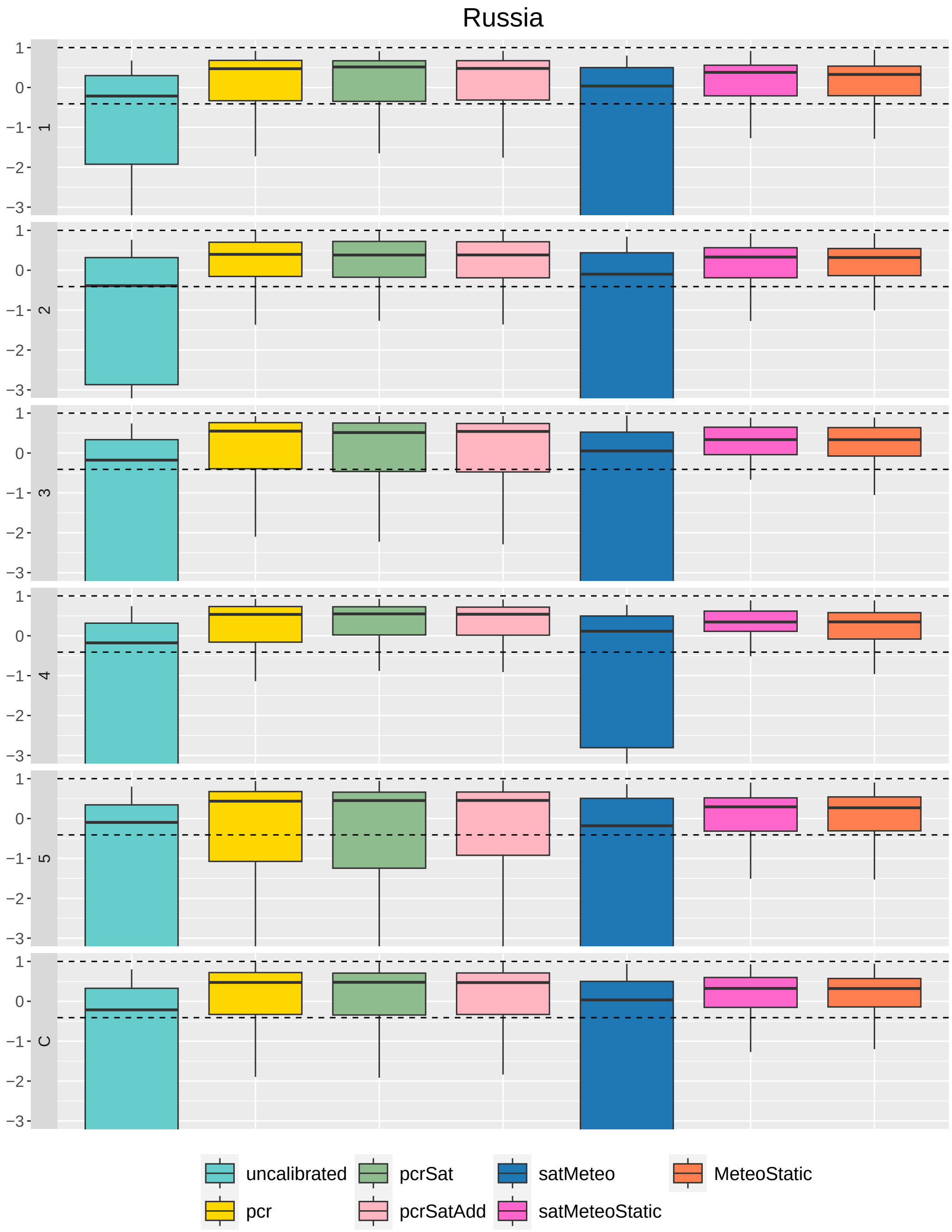


Figure D.4: Boxplots of KGE for all five subsamples and their accumulation for six different configurations in Russia, including uncalibrated PCR-GLOBWB discharge simulations. The dashed lines denote ideal (1) and ’good’ (-0.41) KGE values.

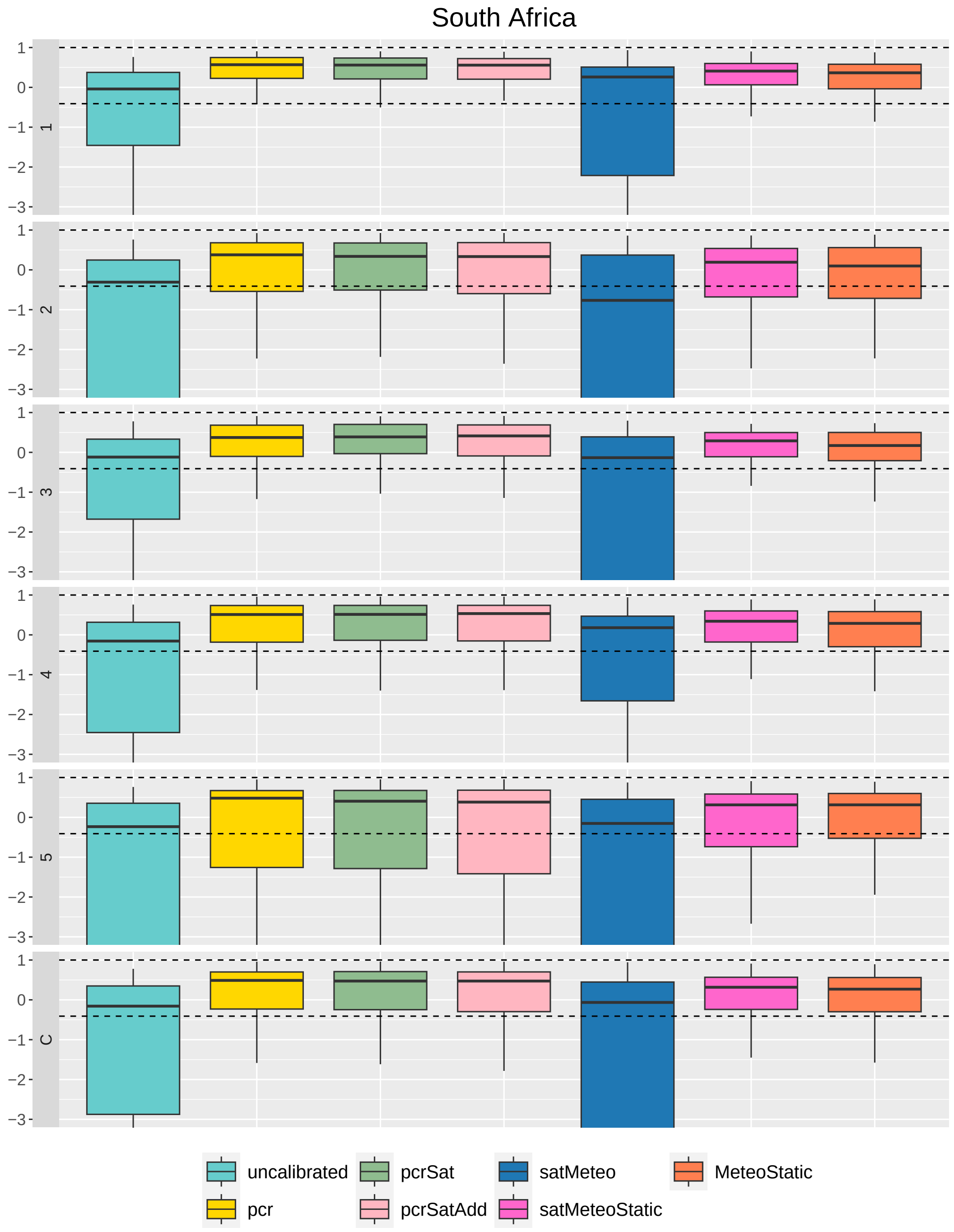


Figure D.5: Boxplots of KGE for all five subsamples and their accumulation for six different configurations in South Africa, including uncalibrated PCR-GLOBWB discharge simulations. The dashed lines denote ideal (1) and ’good’ (-0.41) KGE values.

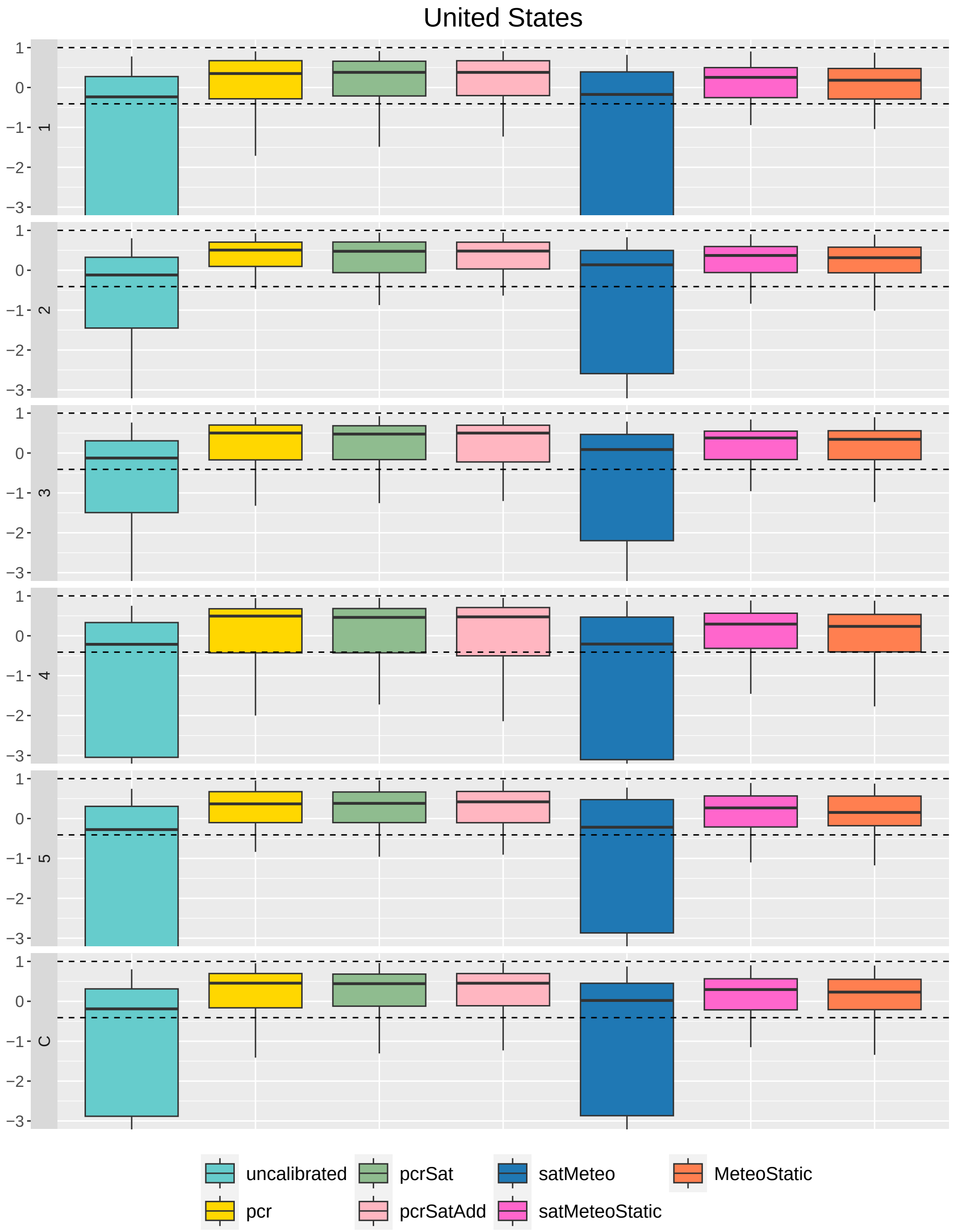


Figure D.6: Boxplots of KGE for all five subsamples and their accumulation for six different configurations in the US, including uncalibrated PCR-GLOBWB discharge simulations. The dashed lines denote ideal (1) and ’good’ (-0.41) KGE values.

D.2 Variable Importance of Local Runs

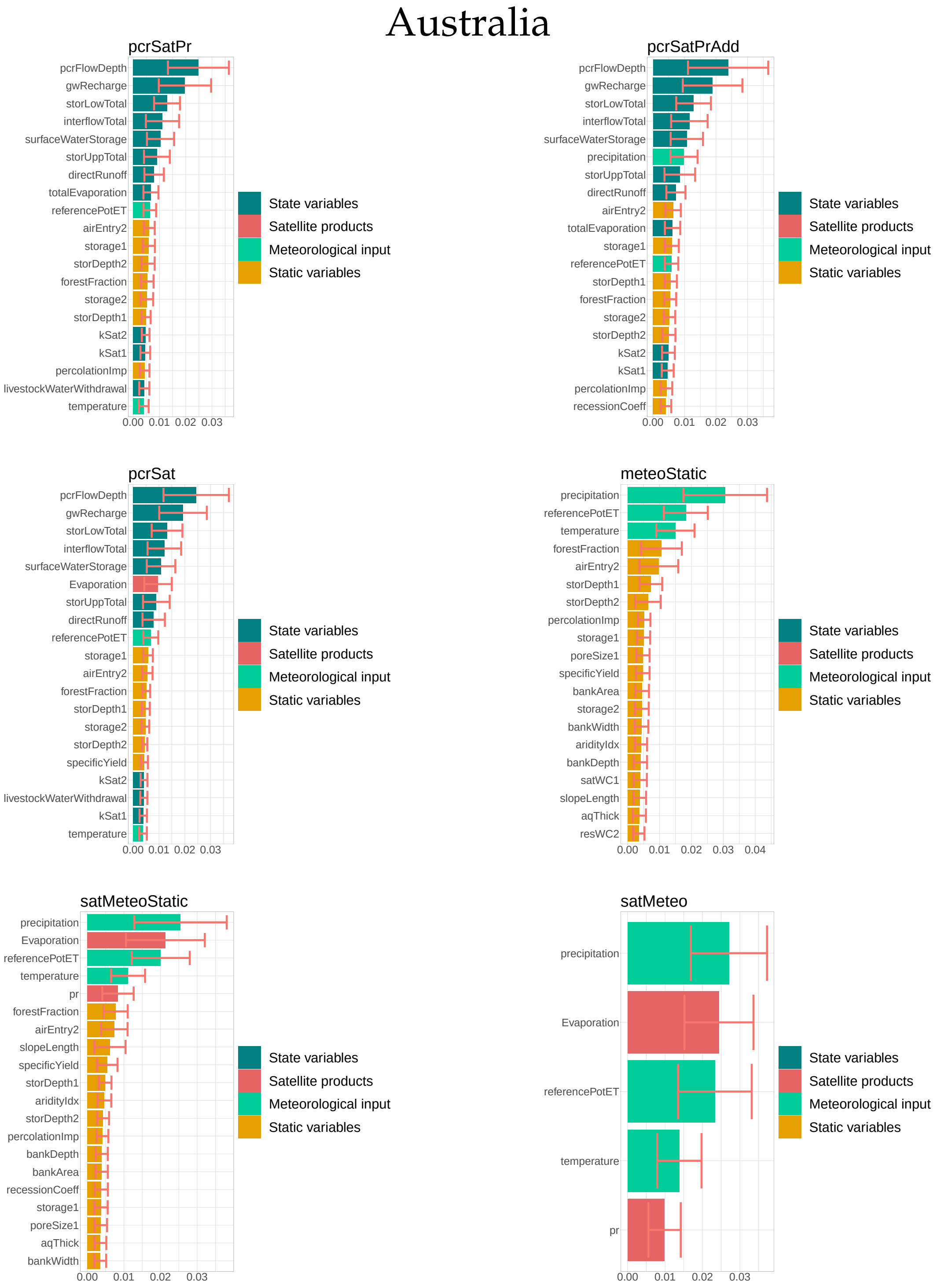


Figure D.7: Square rooted mean decrease in impurity values of the top twenty variables, averaged over five training subsamples for all the Random Forest (RF) configurations for Australia. Each type of variable is represented by a different color.

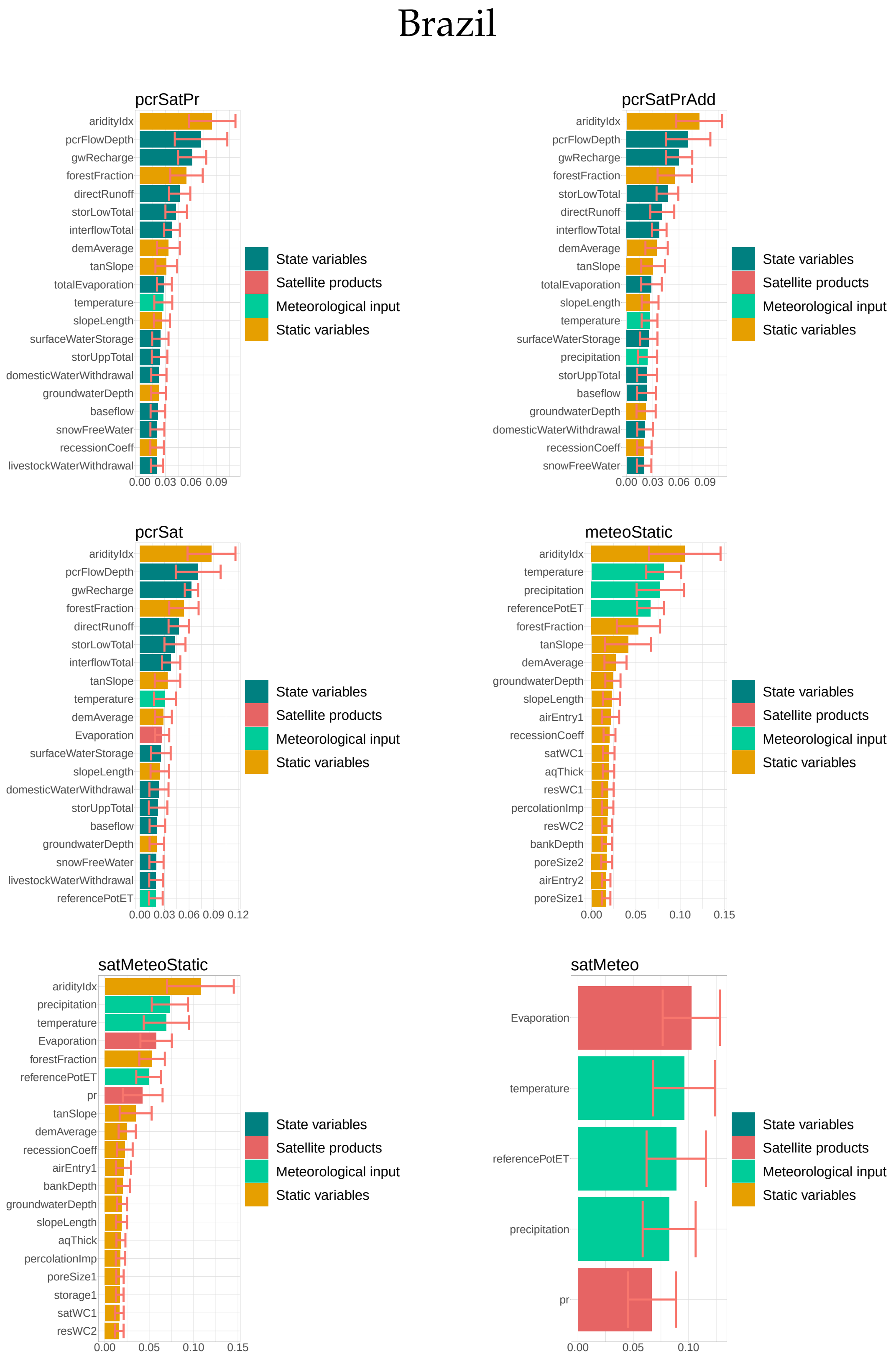


Figure D.8: Square rooted mean decrease in impurity values of the top twenty variables, averaged over five training subsamples for all the Random Forest (RF) configurations for Brazil. Each type of variable is represented by a different color.

Canada

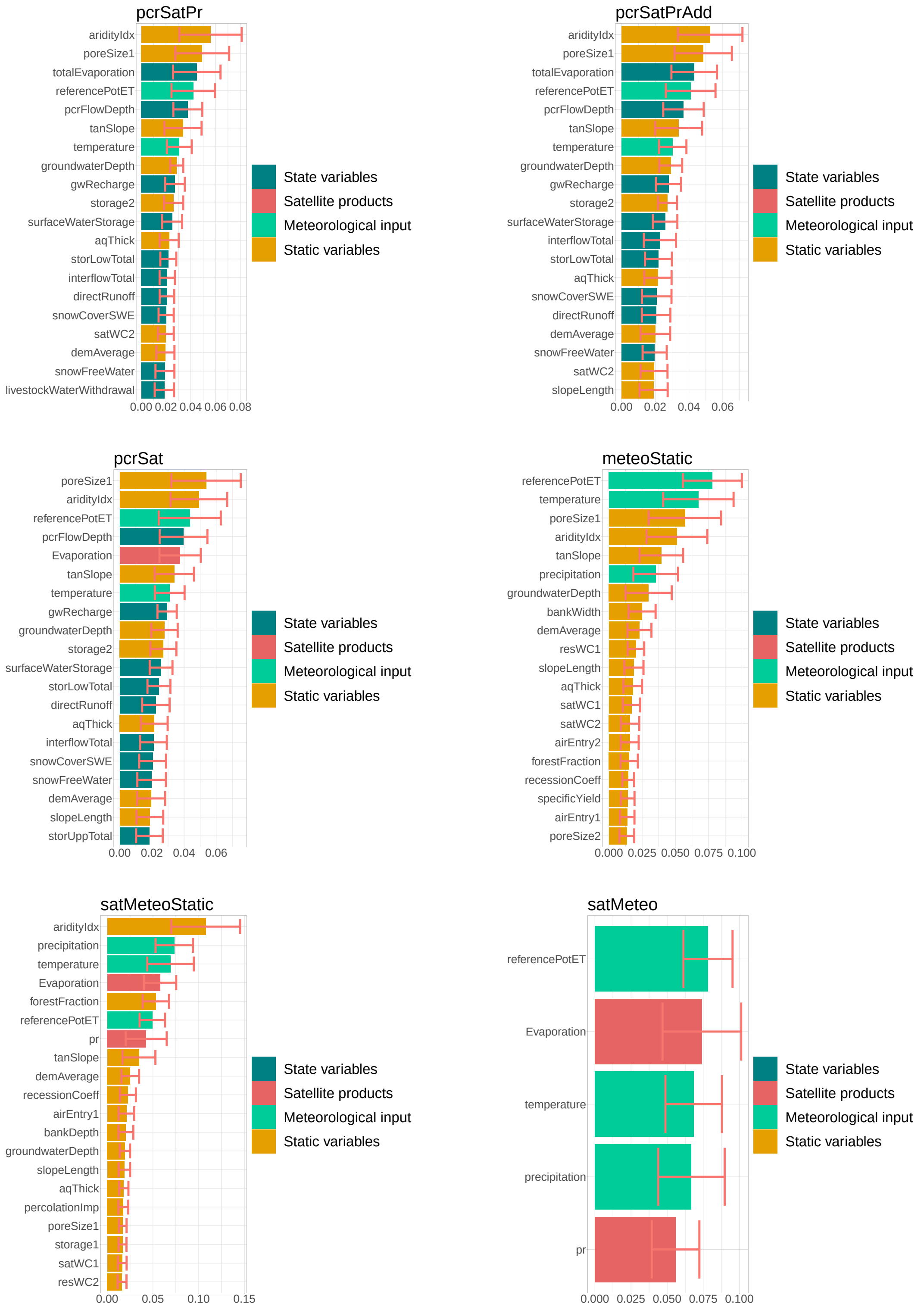


Figure D.9: Square rooted mean decrease in impurity values of the top twenty variables, averaged over five training subsamples for all the Random Forest (RF) configurations for Canada. Each type of variable is represented by a different color.

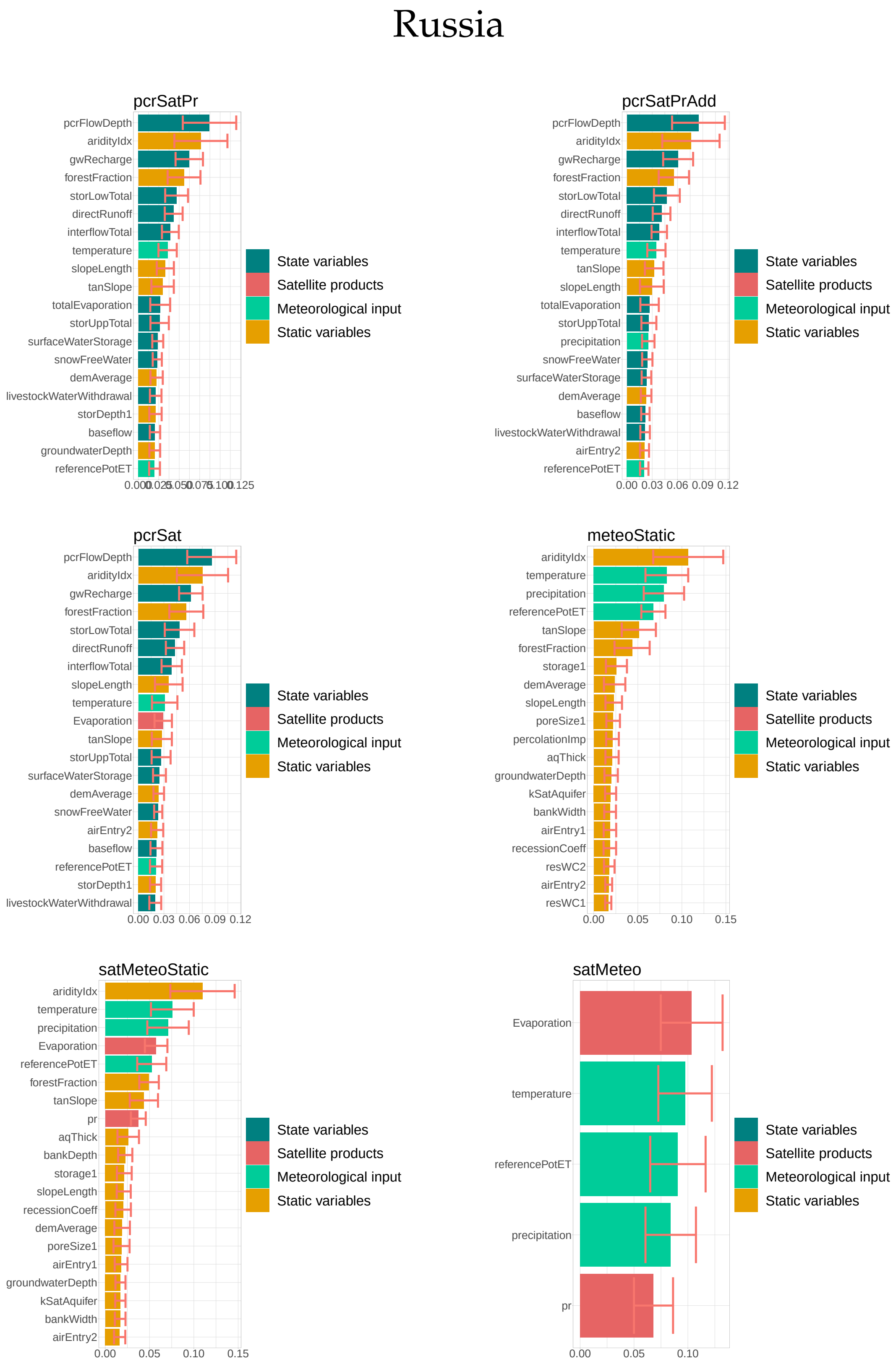


Figure D.10: Square rooted mean decrease in impurity values of the top twenty variables, averaged over five training subsamples for all the Random Forest (RF) configurations for Russia. Each type of variable is represented by a different color.

South Africa

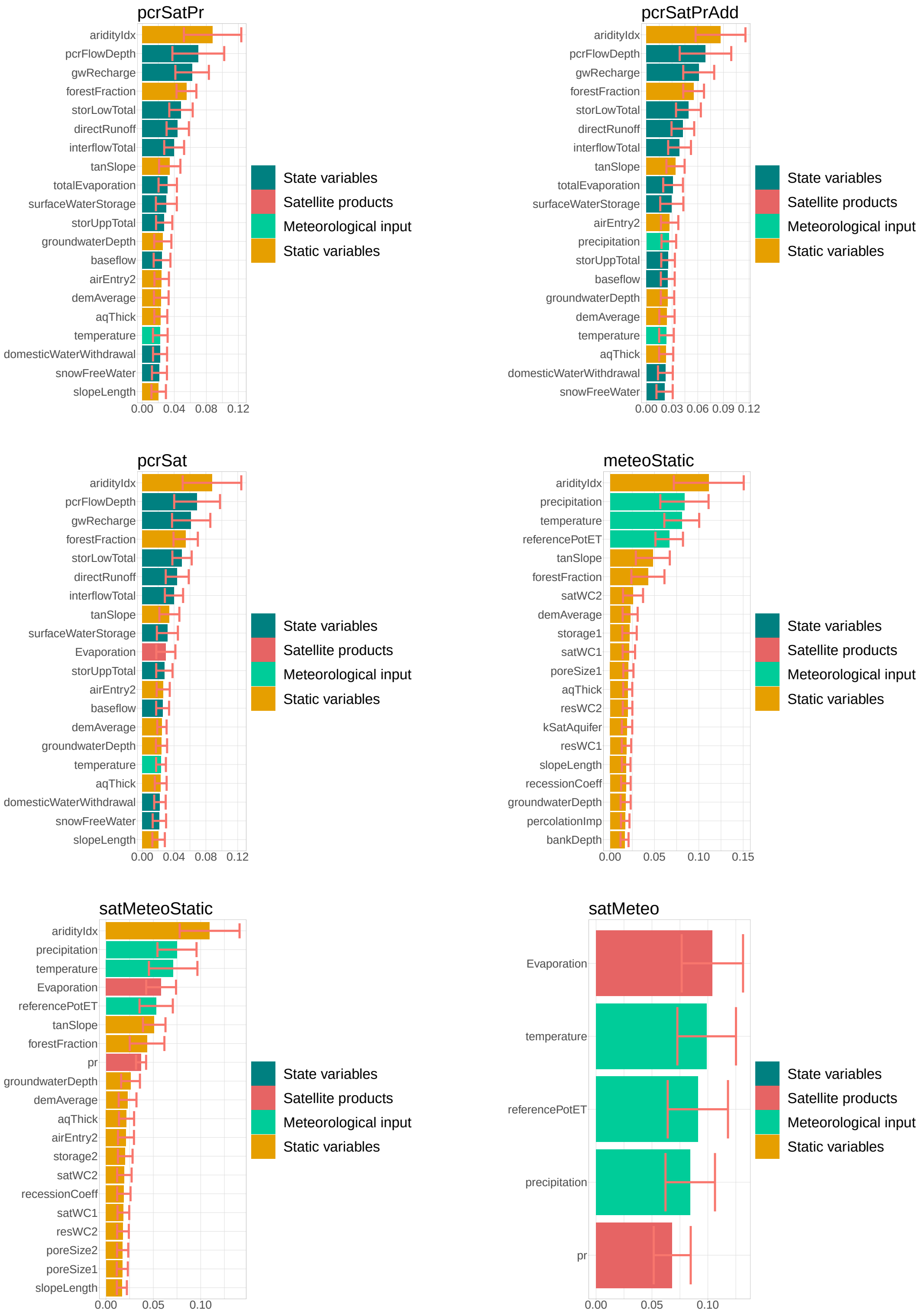


Figure D.11: Square rooted mean decrease in impurity values of the top twenty variables, averaged over five training subsamples for all the Random Forest (RF) configurations for South Africa. Each type of variable is represented by a different color.

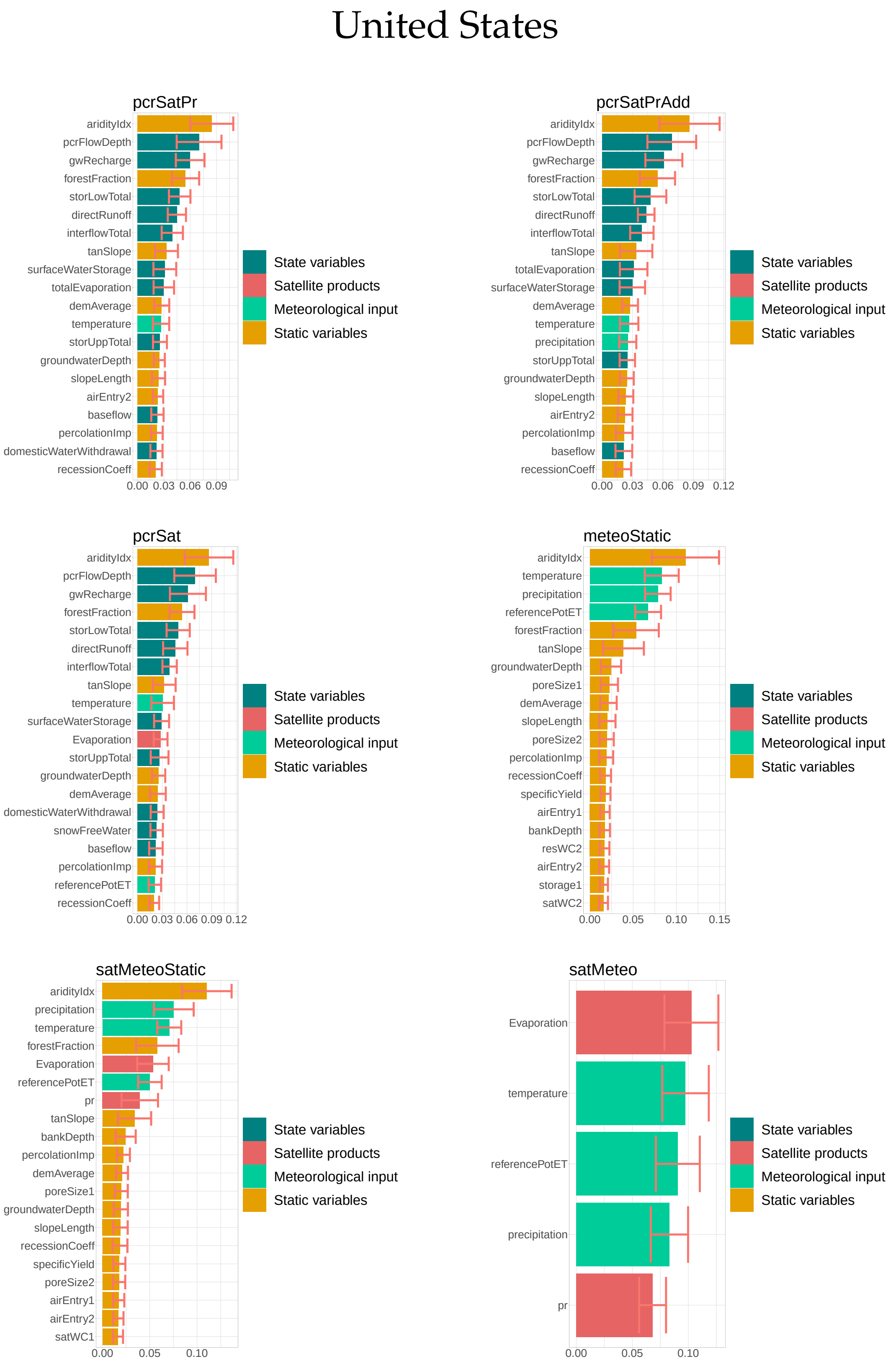


Figure D.12: Square rooted mean decrease in impurity values of the top twenty variables, averaged over five training subsamples for all the Random Forest (RF) configurations for the United States. Each type of variable is represented by a different color.

Bibliography

[1] E. National Academies of Sciences and Medicine, *Attribution of Extreme Weather Events in the Context of Climate Change*. Washington, DC: The National Academies Press, 2016, ISBN: 978-0-309-38094-2. DOI: 10.17226/21852. [Online]. Available: https://nap.nationalacademies.org/catalog/21852/attribution-of-extreme-weather-events-in-the-context-of-climate-change.

[2] K. S. M. H. Ibrahim, Y. F. Huang, A. N. Ahmed, C. H. Koo, and A. El-Shafie, "A review of the hybrid artificial intelligence and optimization modelling of hydrological streamflow forecasting," *Alexandria Engineering Journal*, vol. 61, no. 1, pp. 279–303, 2022, ISSN: 1110-0168. DOI: https://doi.org/10.1016/j.aej.2021.04.100. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S111001682100346X.

[3] F. Fentaw, A. M. Melesse, D. Hailu, and A. Nigussie, "Chapter 10 - precipitation and streamflow variability in tekeze river basin, ethiopia," in *Extreme Hydrology and Climate Variability*, A. M. Melesse, W. Abtew, and G. Senay, Eds., Elsevier, 2019, pp. 103–121, ISBN: 978-0-12-815998-9. DOI: https://doi.org/10.1016/B978-0-12-815998-9.00010-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128159989000105.

[4] L. Alfieri, F. Avanzi, F. Delogu, *et al.*, "High-resolution satellite products improve hydrological modeling in northern italy," *Hydrology and Earth System Sciences*, vol. 26, no. 14, pp. 3921–3939, 2022. DOI: 10.5194/hess-26-3921-2022. [Online]. Available: https://hess.copernicus.org/articles/26/3921/2022/.

[5] P. Sharma and D. Machiwal, *Advances in Streamflow Forecasting - From Traditional to Modern Approaches*. Jun. 2021, ISBN: 9780128206737. DOI: 10.1016/C2019-0-02163-2.

[6] Y. Lin, D. Wang, G. Wang, *et al.*, "A hybrid deep learning algorithm and its application to streamflow prediction," *Journal of Hydrology*, vol. 601, p. 126 636, 2021, ISSN: 0022-1694. DOI: https://doi.org/10.1016/j.jhydrol.2021.126636. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022169421006843.

[7] D. Solomatine and T. Wagener, "2.16 - hydrological modeling," in Dec. 2011, pp. 435–457, ISBN: 9780444531995. DOI: 10.1016/B978-0-444-53199-5.00044-0.

[8] K. Hakala, N. Addor, C. Teutschbein, M. Vis, H. Dakhlaoui, and J. Seibert, *Hydrological modeling of climate change impacts*, Dec. 2019. DOI: 10.1002/9781119300762.wsts0062. [Online]. Available: http://dx.doi.org/10.1002/9781119300762.wsts0062.

[9] R. Merz, J. Parajka, and G. Blöschl, "Time stability of catchment model parameters: Implications for climate impact analyses," *Water Resources Research*, vol. 47, no. 2, Feb. 2011, ISSN: 1944-7973. DOI: 10.1029/2010wr009505. [Online]. Available: http://dx.doi.org/10.1029/2010WR009505.

[10] H. Dakhlaoui, D. Ruelland, Y. Tramblay, and Z. Bargaoui, "Evaluating the robustness of conceptual rainfall-runoff models under climate variability in northern tunisia," *Journal of Hydrology*, vol. 550, pp. 201–217, Jul. 2017, ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2017.04.032. [Online]. Available: http://dx.doi.org/10.1016/j.jhydrol.2017.04.032.

[11] F. Kratzert, D. Klotz, M. Herrnegger, A. K. Sampson, S. Hochreiter, and G. S. Nearing, "Toward improved predictions in ungauged basins: Exploiting the power of machine learning," *Water Resources Research*, vol. 55, no. 12, pp. 11 344–11 354, Dec. 2019, ISSN: 1944-7973. DOI: 10.1029/2019wr026065. [Online]. Available: http://dx.doi.org/10.1029/2019WR026065.

[12] M. Ghaith, A. Siam, Z. Li, and W. El-Dakhakhni, "Hybrid hydrological data-driven approach for daily streamflow forecasting," *Journal of Hydrologic Engineering*, vol. 25, no. 2, Feb. 2020, ISSN: 1943-5584. DOI: 10.1061/(asce)he.1943-5584.0001866. [Online]. Available: http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0001866.

[13] B. Kraft, M. Jung, M. Körner, S. Koirala, and M. Reichstein, "Towards hybrid modeling of the global hydrological cycle," *Hydrology and Earth System Sciences*, vol. 26, no. 6, pp. 1579–1614, 2022. DOI: 10.5194/hess-26-1579-2022. [Online]. Available: https://hess.copernicus.org/articles/26/1579/2022/.

[14] R. Remesan and J. Mathew, "Data-based evapotranspiration modeling," in *Hydrological Data Driven Modelling: A Case Study Approach*. Cham: Springer International Publishing, 2015, pp. 183–230, ISBN: 978-3-319-09235-5. DOI: 10.1007/978-3-319-09235-5_7. [Online]. Available: https://doi.org/10.1007/978-3-319-09235-5_7.

[15] L. J. Slater, L. Arnal, M.-A. Boucher, *et al.*, "Hybrid forecasting: Blending climate predictions with ai models," *Hydrology and Earth System Sciences*, vol. 27, no. 9, pp. 1865–1889, 2023. DOI: 10.5194/hess-27-1865-2023. [Online]. Available: https://hess.copernicus.org/articles/27/1865/2023/.

[16] H. Tyralis, G. Papacharalampous, and A. Langousis, "A brief review of random forests for water scientists and practitioners and their recent history in water resources," *Water*, vol. 11, p. 910, Apr. 2019. DOI: 10.3390/w11050910.

[17] Y. Shen, J. Ruijsch, M. Lu, E. H. Sutanudjaja, and D. Karssenberg, "Random forests-based error-correction of streamflow from a large-scale hydrological model: Using model state variables to estimate error terms," *Computers Geosciences*, vol. 159, p. 105 019, 2022, ISSN: 0098-3004. DOI: https://doi.org/10.1016/j.cageo.2021.105019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0098300421003010.

[18] E. H. Sutanudjaja, R. van Beek, N. Wanders, *et al.*, "PCR-GLOBWB 2: A 5 arcmin global hydrological and water resources model," *Geoscientific Model Development*, vol. 11, no. 6, pp. 2429–2453, 2018. DOI:

10.5194/gmd-11-2429-2018. [Online]. Available: https://gmd.copernicus.org/articles/11/2429/2018/.

[19] M. Magni, E. Sutanudjaja, Y. Shen, and D. Karssenberg, "Global streamflow modelling using process-informed machine learning," *Journal of Hydroinformatics*, Aug. 2023. DOI: 10.2166/hydro.2023.217.

[20] N. C. d'Escury, *Using satellite products to improve random forest-based streamflow simulations from a global hydrological model*, Jun. 2023.

[21] D. G. Miralles, T. R. H. Holmes, R. A. M. De Jeu, J. H. Gash, A. G. C. A. Meesters, and A. J. Dolman, "Global land-surface evaporation estimated from satellite-based observations," *Hydrology and Earth System Sciences*, vol. 15, no. 2, pp. 453–469, 2011. DOI: 10.5194/hess-15-453-2011. [Online]. Available: https://hess.copernicus.org/articles/15/453/2011/.

[22] S. Zhu, J. Wei, H. Zhang, Y. Xu, and H. Qin, "Spatiotemporal deep learning rainfall-runoff forecasting combined with remote sensing precipitation products in large scale basins," *Journal of Hydrology*, vol. 616, p. 128 727, Jan. 2023, ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2022.128727. [Online]. Available: http://dx.doi.org/10.1016/j.jhydrol.2022.128727.

[23] G. J. Huffman, D. T. Bolvin, D. Braithwaite, *et al.*, "Integrated multi-satellite retrievals for the global precipitation measurement (gpm) mission (imerg)," in *Satellite Precipitation Measurement*. Springer International Publishing, 2020, pp. 343–353, ISBN: 9783030245689. DOI: 10.1007/978-3-030-24568-9_19. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-24568-9_19.

[24] M. Magni, E. H. Sutanudjaja, Youchen Shen, and D. Karssenberg, *Input data for PCR-GLOBWB-RF (30 arcmin)*, 2023. DOI: 10.5281/ZENODO.7890583. [Online]. Available: https://zenodo.org/record/7890583.

[25] B. Martens, D. G. Miralles, H. Lievens, *et al.*, "GLEAM v3: Satellite-based land evaporation and
root-zone soil moisture," *Geoscientific Model Development*, vol. 10, no. 5, pp. 1903–1925, 2017. DOI: 10.5194/gmd-10-1903-2017. [Online]. Available: https://gmd.copernicus.org/articles/10/1903/2017/.

[26] Z. Wang, R. Zhong, C. Lai, and J. Chen, "Evaluation of the GPM IMERG satellite-based precipitation products and the hydrological utility," *Atmospheric Research*, vol. 196, pp. 151–163, Nov. 2017, ISSN: 0169-8095. DOI: 10.1016/j.atmosres.2017.06.020. [Online]. Available: http://dx.doi.org/10.1016/j.atmosres.2017.06.020.

[27] N. G. P. M. Mission, *IMERG Final Run*, Accessed: 2024. [Online]. Available: https://gpm.nasa.gov/taxonomy/term/1417#:~:text=The%20IMERG%20dataset%20now%20includes,are%20considered%20%22partial%20coverage%22..

[28] U. Schulzweida, "CDO User Guide," en, 2022. DOI: 10.5281/ZENODO.7112925. [Online]. Available: https://zenodo.org/record/7112925.

[29] D. Karssenberg, O. Schmitz, P. Salamon, K. de Jong, and M. F. Bierkens, "A software framework for construction of process-based stochastic spatio-temporal models and data assimilation," *Environmental Modelling & Software*, vol. 25, no. 4, pp. 489–502, 2010, ISSN: 1364-8152. DOI: https://doi.org/10.1016/j.envsoft.2009.10.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364815209002643.

[30] S. Lange, C. Menz, S. Gleixner, *et al.*, *WFDE5 over land merged with ERA5 over the ocean (W5E5 v2.0)*, en, 2021. DOI: 10.48364/ISIMIP.342217. [Online]. Available: https://data.isimip.org/10.48364/ISIMIP.342217.

[31] L. Breiman, *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, ISSN: 0885-6125. DOI: 10.1023/a:1010933404324. [Online]. Available: http://dx.doi.org/10.1023/A:1010933404324.

[32] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *WIREs Data Mining and Knowledge Discovery*, vol. 2, no. 6, pp. 493–507, Oct. 2012, ISSN: 1942-4795. DOI: 10.1002/widm.1072. [Online]. Available: http://dx.doi.org/10.1002/widm.1072.

[33] In *Encyclopedia of Machine Learning*. Springer US, 2011, pp. 828–828, ISBN: 9780387301648. DOI: 10.1007/978-0-387-30164-8_695. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-30164-8_695.

[34] P. Probst, M. N. Wright, and A.-L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 3, Jan. 2019, ISSN: 1942-4795. DOI: 10.1002/widm.1301. [Online]. Available: http://dx.doi.org/10.1002/widm.1301.

[35] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, Apr. 2016, ISSN: 1863-8260. DOI: 10.1007/s11749-016-0481-7. [Online]. Available: http://dx.doi.org/10.1007/s11749-016-0481-7.

[36] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, Jan. 2013. DOI: 10.1007/978-1-4614-7138-7. [Online]. Available: https://doi.org/10.1007/978-1-4614-7138-7.

[37] M. N. Wright and A. Ziegler, "Ranger: A fast implementation of random forests for high dimensional data in C++ and R," *Journal of Statistical Software*, vol. 77, no. 1, 2017, ISSN: 1548-7660. DOI: 10.18637/jss.v077.i01. [Online]. Available: http://dx.doi.org/10.18637/jss.v077.i01.

[38] H. V. Gupta, H. Kling, K. K. Yilmaz, and G. F. Martinez, "Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling," *Journal of Hydrology*, vol. 377, no. 1, pp. 80–91, 2009, ISSN: 0022-1694. DOI: `https://doi.org/10.1016/j.jhydrol.2009.08.003`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0022169409004843`.

[39] *Mwangi, P., Okelo, O.W., Kamande, K.F. and Mwende, M.J. (2020) A Climate-Smart Agriculture Approach Using Double Digging, Zai Pits and Aquacrop Model in Rain-Fed Sorghum Cultivation at Wiyumiririe Location of Laikipia County, Kenya. Africa Journal of Physical Sciences, 4, 23-53. - References - Scientific Research Publishing — scirp.org*, `https://www.scirp.org/reference/referencespapers?referenceid=2911317`, [Accessed 26-02-2024].

[40] D. Jiang and K. Wang, "The Role of Satellite-Based Remote Sensing in Improving Simulated Streamflow: A Review," *Water*, vol. 11, no. 8, 2019, ISSN: 2073-4441. DOI: `10.3390/w11081615`. [Online]. Available: `https://www.mdpi.com/2073-4441/11/8/1615`.

[41] C. Deval, "Integration of remote sensing data on precipitation, evapotranspiration leaf area index into the distributed global hydrological model PCR-GLOBWB," Ph.D. dissertation, Aug. 2016. DOI: `10.13140/RG.2.2.20080.97280`.

[42] K. Ma, D. Feng, K. Lawson, *et al.*, "Transferring Hydrologic Data Across Continents – Leveraging Data-Rich Regions to Improve Hydrologic Prediction in Data-Sparse Regions," *Water Resources Research*, vol. 57, no. 5, Apr. 2021, ISSN: 1944-7973. DOI: `10.1029/2020wr028600`. [Online]. Available: `http://dx.doi.org/10.1029/2020WR028600`.