Title
Improved Hydrologic Forecasting and Hydropower Planning In Data Scarce Regions Using Satellite-Based Remote Sensing

Author
Koppa, Akash

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Improved Hydrologic Forecasting and Hydropower Planning In Data Scarce Regions Using Satellite-Based Remote Sensing

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Civil Engineering

by

Akash Koppa

2019

ABSTRACT OF THE DISSERTATION

Improved Hydrologic Forecasting and Hydropower Planning In Data Scarce Regions Using Satellite-Based Remote Sensing

by

Akash Koppa
Doctor of Philosophy in Civil Engineering
University of California, Los Angeles, 2019
Professor Mekonnen Gebremichael, Chair

The role of satellite-based remote sensing in improving hydrologic and water resources studies in data-scarce regions is investigated. Specifically, the dissertation focuses on the development of a 1) validation framework for remotely sensed precipitation and evapotranspiration without the use of ground-based observations, 2) methodological framework for calibration of large scale hydrologic models with multiple fluxes, and 3) a seasonal hydropower planning framework for data-scarce regions. In the first part of the dissertation, a root mean square error (RMSE)-based error metric capable of translating individual biasies in precipitation and evapotranspiration onto the Budyko space is developed. It is shown that the framework succeeds in arriving at the same conclusions as a traditional validation method. In the second part, the value of incorporating multiple hydrologic variables such as evapotranspiration, soil moisture and streamflow into model calibration is investigated. It is shown that parameters which are insensitive to individual model responses can influence the trade-off relationship between them. Finally, the potential of using remotely sensed precipitation and evapotranspiration datasets in generating reliable seasonal reservoir inflow forecasts for hydropower planning is investigated. Results highlight the importance of accounting for input and parameter uncertainty in hydropower planning.

The dissertation of Akash Koppa is approved.

William W-G Yeh

Steven A. Margulis

Timu W. Gallien

Mekonnen Gebremichael, Committee Chair

University of California, Los Angeles

2019

To my parents

TABLE OF CONTENTS

List of Figures . . . . . **viii**

List of Tables . . . . . **xiii**

Acknowledgments . . . . . **xv**

Vita . . . . . **xvi**

1 Introduction **1**

1.1 Background and motivation 1

1.2 Seasonal hydropower planning in data-scarce regions 3

1.2.1 Climate system 4

1.2.2 Hydrology system 6

1.2.3 Optimization system 7

1.3 Organization of the dissertation 8

2 A Framework for Validation of Remotely Sensed Precipitation and Evapotranspiration Based on the Budyko Hypothesis **9**

2.1 Introduction 9

2.2 Study area and observational data 12

2.3 Development of the validation framework 12

2.3.1 Budyko Space 12

2.3.2 Defining the error metric 13

2.3.3 Estimating the error metric 14

2.3.4 Sensitivity analysis of the error metric 15

2.3.5 Modification of the error metric . . . 17
2.3.6 Sensitivity of modified distance to ω and AI . . . 18
2.3.7 Statistical significance test . . . 19
2.4 Validating the framework . . . 20
2.4.1 Validation in the Budyko space . . . 21
2.4.2 Validation with traditional RMSE . . . 24
2.4.3 Statistical significance of the results . . . 25
2.5 Applying the framework to a data-scarce catchment . . . 26
2.6 Conclusions, limitations and future work . . . 28

3 Calibration of Large Scale Hydrologic Models with Multiple Fluxes: The Necessity and Value of a Pareto Optimal Approach . . . 30

3.1 Introduction . . . 30
3.2 Methodology . . . 34
3.2.1 Conceptual framework . . . 34
3.2.2 Defining limits of acceptability for individual model responses . . . 36
3.2.3 Pareto optimal solutions for combination of model responses . . . 37
3.2.4 Hypothesis testing, trade-off analysis, and model diagnosis . . . 38
3.3 Experiment design . . . 40
3.3.1 Study area and time period . . . 40
3.3.2 Observational data . . . 40
3.3.3 Setup and validation of the hydrologic model . . . 43
3.3.4 Setup of DREAM and AMALGAM algorithms . . . 47
3.4 Results and discussions . . . 51
3.4.1 Posterior distributions of model errors . . . 51

3.4.2 Hypothesis testing using DREAM and AMALGAM solutions 55
3.4.3 Understanding the trade-offs using Pareto fronts 57
3.4.4 Model diagnosis 60
3.5 Conclusions and future work 67

4 Seasonal Hydropower Planning For Data Scarce Regions Using Multi Model Ensemble Forecasts, Remote Sensing Data, and Stochastic Programming . 70

4.1 Introduction 70
4.2 Methodology 73
4.3 Experiment design 78
4.3.1 Study area and time period 78
4.3.2 Observational and forecast data 79
4.3.3 Hydrologic model: setup, calibration and model parameter uncertainty 81
4.3.4 Hydropower optimization model 83
4.4 Results and discussion 84
4.4.1 Validation of the NMME precipitation forecasts 84
4.4.2 Validation of the Noah-MP hydrologic model 88
4.4.3 Seasonal Hydropower planning in the study region 90
4.5 Conclusions and future work 98

5 Dissertation Conclusions and Future Work 100

5.1 Conclusions and original contributions 100
5.2 Future work 102

LIST OF FIGURES

1.1 Spatial and temporal distribution of streamflow gage stations available in the Global Runoff Data Centre (sourced from https://www.bafg.de/). 2

1.2 A representation of a seasonal hydropower planning system consisting of the climate, hydrology and optimization components and their linkage. 4

2.1 a) MOPEX catchments in the United States classified according to aridity and b) A representation of the Budyko space consisting of Fu's curve with $\omega = 2.6$ (Fu, 1981), water and energy limits. The original definition of distance following Greve et al. (2014) and the modified distances (MDm and MDo) are represented. Four cases of estimated (AI, EI) points representing different precipitation biases are shown: Case 1 - Low positive bias, Case 2 - High positive bias, Case 3 - Low negative bias and Case 4 - High negative bias (Refer to section 3.4). 11

2.2 Sensitivity maps of the logarithm of 1) RMSE_{OD} (Original Distance) and 2) RMSE_{MDm} (Modified distance). The red star represents RMSE value for no bias in either precipitation or evapotranspiration. n is number of catchments in each hydroclimatic regime. 15

2.3 Bar Plots of RMSE_{MDm} and RMSE_{MDo} values for different combinations of P and E datasets and for all hydroclimatic regimes. The three panels correspond to the three satellite-based Precipitation products. 22

2.4 Scatter plots of observed modified distance (X-axis) versus modeled modified distance (Y-axis) for the MOPEX catchments and for all combinations of remotely sensed P and E datasets. 23

2.5 Bar Plots of RMSE_{T} (Traditional RMSE) values for different P and E datasets and for all hydroclimatic regimes. 24

3.1 Conceptual framework of the methodology adopted in this study (adapted from Efstratiadis and Koutsoyiannis (2010)). 36

3.2 A map of the Mississippi basin showing the six USGS HUC-2 basins. 41

3.3 Scatter plots of a) GLEAM ET vs Ameriflux measurements (top right panel); b) ESA-CCI soil moisture vs TAMU NASMDB measurements (bottom panel) for 2004; c) annual GLEAM ET vs Annual Precipitation (P) - Runoff (Q) for the years 2000-2009 (bottom left). The errors for the calibration and validation years (2004 and 2005) are highlighted by black and brown bounding boxes respectively; and d) the Budyko space (Evaporative index vs Aridity index) averaged over the years 2000-2009 for the six catchments. The red line is the ideal catchment water-energy balance as represented by the Budyko hypothesis. The dotted lines represent the water (horizontal line) and energy (diagonal line) limits (bottom right). 44

3.4 a) Time series plots of a) ET-calibrated model ET (red) and GLEAM observed ET (black) in mm/month (Top panel), b) SM-calibrated model SM (red) and ESA-CCI observed SM (black) in m^3/m^3 (Middle panel) and c) SF-calibrated model SF (red) and observed HUC-2 runoff (black) in mm/month (Bottom panel) for the six HUC-2 sub-catchments of the Mississippi basin. The first 12 months correspond to the calibration period of 2004 and the next 24 months correspond to validation years 2005; and b) Scatter plot of annual modeled ET, SM and SF and observed values for the calibration and validation years over the six HUC-2 basins. 49

3.5 A comparison of the posterior probability density functions (PDF) and empirical cumulative distribution functions (ECDF) of root mean square errors of a) evapotranspiration (top panel), b) soil moisture (middle panel) and c) streamflow (bottom panel) when the model is calibrated using DREAM with ET (green), SM (orange) and SF (blue). Vertical lines in the PDFs represent 50% quantiles of RMSE. For comparison, the mean annual water balance closure error from the observational datasets (P - Q - ET) is around 108 mm (9 mm/month). Note: The RMSE values of ET and SM are determined from error residuals calculated at all grid cells of the model domain. The RMSE values of SF are determined using error residuals from all the six HUC-2 regions. 53

3.6 Pareto fronts of root mean square errors of a) ET and SM, b) ET and SF and c) SM and SF with limits of acceptability represented by 10%, 25%, 50%, 75%, 90%, 95% and 99% quantiles of the posterior distribution of the error (from DREAM). For comparison, the mean annual water balance closure error from the observational datasets (P - Q - ET) is around 108 mm (9 mm/month). The non-dominated or Pareto optimal solutions are represented by red stars. Note: The RMSE values of ET and SM are determined from error residuals calculated at all grid cells of the model domain. The RMSE values of SF are determined using error residuals from all the six HUC-2 regions. 54

3.7 Empirical cumulative density functions (ECDF) of calibrated Noah-MP parameters for univariate (ET, SM, and SF) and multivariate (ET-SM, ET-SF, and SM-SF) objectives. For the ET, SM, SF, and ET-SF objectives, the solutions within the error threshold of the 50% quantile are used to construct the PDFs and ECDFs of the parameters. For the ET-SM and SM-SF solutions that are not behavioral for both, model responses at the 50% quantile error threshold are used. 65

3.8 a) Hellingers distance between PDFs of the calibration parameters for different calibration objectives and b) the Kolmogorov-Smirnov test statistic between the ECDFs of calibration parameters for dierent calibration objectives. For the ET, SM, SF and ET-SF objectives, the solutions within the error threshold of the 50% quantile are used to construct the PDFs and ECDFs of the parameters. For ET-SM and SM-SF solutions that are not behavioral for both, the model responses at the 50% quantile error threshold are used. 66

4.1 A visual representation of three scenario fan structures A) A single deterministic forecast (DET), B) first stage deterministic and the rest stochastic (SPWR-D) and C) all stages stochastic (SPWR-S). 77

4.2 Flow chart for seasonal hydropower planning using a) DET and SPWR-D and b) SPWR-S reservoir inflow scenario structures. The precipitation forecasts, observational datasets, hydrologic model, and optimization algorithm used in the case study are mentioned in parenthesis. 78

4.3 The Omo-Gibe river basin consisting of a cascade of five reservoirs, located in East Africa. The countries presented constitute the East African Power Pool (EAPP). 79

4.4 Time series comparison of precipitation from different NMME models and observations (TRMM) for different lead times (1-8 months) and for the calibration (12 months of 2004) and validation (12 months of 2005) time periods. We present the mean of the 10 members from CanCM3, CanCM4 and GEOS-5. Ensemble Mean and Ensemble BMA are the simple mean and Bayesian model average of all the 30 ensemble members. In addition, we present the 5% and 95% quantiles of all the 30 ensemble members. 86

4.5 Taylor diagrams of precipitation from CanCM3, CanCM4, GEOS-5, Ensemble Mean and BMA models determined for different lead times (1-8 months) and for the calibration (12 months of 2004) and validation (12 months of 2005) time periods. 87

4.6 Time series comparison of evapotranspiration from ET-calibrated Noah-MP model (green) and observed ET estimates (black) from GLEAM for the calibration (12 months of 2004) and validation (12 months of 2005) time periods. The 5% and 95% quantiles from the behavioral solutions of Bayesian calibration is used to determine the uncertainty in modeled ET (green band). 89

4.7 Posterior probability density functions (PDFs) of the Noah-MP hydrologic model parameters considered for calibration using the MT-DREAM (ZS) algorithm and ET estimates from GLEAM. The green line represents the 50% quantile values for each of the parameters. 90

4.8 a) Reservoir inflows, b) optimal release decisions, c) storage and d) power produced in the Gibe 1 and Gibe 3 reservoirs for the 8 month planning horizon (February 2005 to September 2005) and three scenario structures (DET, SPWR-D and SPWR-S). Note: the results correspond to the first stage of each iteration of the deterministic and stochastic programming with recourse model. 94

4.9 Uncertainty in a) reservoir inflows, b) optimal release decisions, c) storage and d) power produced in the Gibe 1 and Gibe 3 reservoirs for the 8 month planning horizon (February 2005 to September 2005) and three scenario structures (DET, SPWR-D and SPWR-S), due to uncertainty in model parameters derived from Bayesian calibration. Note: the results correspond to the first stage of each iteration of the deterministic and stochastic programming with recourse model. . 95

LIST OF TABLES

2.1 Variance and mean (in brackets) of the modified distance (MDm) for 1) Fixed AI_{act} and varying ω and 2) Fixed ω and varying AI_{act}. Depending on the values of Estimated (AI_{est}, EI_{est}) and actual AI_{act}, 18 cases are presented. The fixed AI_{act} values for humid, temperate and arid regions are 0.5, 1.5 and 4 respectively. The fixed ω value is 2.6 20

3.1 Noah-MP model physics options 47

3.2 Details of Noah-MP parameters for calibration 48

3.3 MT-DREAM (ZS) and AMALGAM configuration 50

3.4 Limits of Acceptability for ET, SM and SF derived from posterior distributions of RMSE 55

3.5 Breakdown of Pareto optimal solutions into behavioral solutions based on different limits of acceptability for ET-SM, ET-SF and SM-SF combinations 58

3.6 Average, minimum and maximum trade-offs for combinations of ET, SM and SF model responses (minimum and maximum values are within parentheses) 60

3.7 Correlation between objective functions (RMSE) and Noah-MP model parameters (REFDK, REFKDT, BB, MAXSMC and SATDK) for three multivariate calibration cases (ET-SM, ET-SF and SM-SF) 67

4.1 Details of the five hydropower reservoirs in the Omo-Gibe river basin 80

4.2 Different measures of absolute and relative dispersion determined for the ensemble inflows into Gibe I and Gibe III reservoirs generated using the best performing Noah-MP parameter set 96

4.3 Different measures of absolute and relative dispersion determined for the ensemble inflows into Gibe I and Gibe III reservoirs generated using all the behavioral Noah-MP parameter sets 97

ACKNOWLEDGMENTS

I wish to acknowledge the support of my advisor Dr. Mekonnen Gebremichael in guiding me through my doctoral research. I am indebted to the guidance of Dr. William Yeh, without whom this dissertation would not be possible. I thank the dissertation committee members Dr. Steve Margulis and Dr. Timu Gallien for their valuable suggestions. I also wish to thank my family and friends for being there through the ups and downs over the last four years. I acknowledge the funding support from NASA's Applied Sciences Program Water Resource Application program (NNX15AC33G).

Chapter 2 contains the following published article:
Koppa, A., and Gebremichael, M. (2017). A framework for validation of remotely sensed precipitation and evapotranspiration based on the Budyko hypothesis. Water Resources Research, 53, 8487 8499. https://doi.org/10.1002/2017WR020593.

Chapter 3 contains the following submitted journal article:
Koppa, A.,Gebremichael, M., and Yeh, W. W.-G. (Under Review). Calibration of large scale hydrologic models with multiple fluxes: The necessity and value of a Pareto optimal approach. Advances in Water Resources.

Chapter 4 contains the following submitted journal article:
Koppa, A.,Gebremichael, M., Zambon, R., Yeh, W. W.-G., Hopson, T. (Under Review). Seasonal Hydropower Planning For Data Scarce Regions Using Multi Model Ensemble Forecasts, Remote Sensing Data, and Stochastic Programming. Water Resources Research.

VITA

2012	B.E. (Civil Engineering), BMS College of Engineering, Bengaluru, India
2014	M.Tech, (Water Management), Indian Institute of Technology Kharagpur, Kharagpur, India
2015-2019	Graduate Student Researcher, Department of Civil and Environmental Engineering, University of California, Los Angeles

PUBLICATIONS

Koppa, A., and Gebremichael, M. (2017). A framework for validation of remotely sensed precipitation and evapotranspiration based on the Budyko hypothesis. Water Resources Research, 53, 8487 8499. https://doi.org/10.1002/2017WR020593.

Koppa, A.,Gebremichael, M., and Yeh, W. W.-G. (Under Review). Calibration of large scale hydrologic models with multiple fluxes: The necessity and value of a Pareto optimal approach. Advances in Water Resources.

Koppa, A.,Gebremichael, M., and Yeh, W. W.-G. (Under Review). Calibration of large scale hydrologic models with multiple fluxes: The necessity and value of a Pareto optimal approach. Advances in Water Resources.

Xiao, M., Koppa, A., Mekonnen, Z., Pagn, B. R., Zhan, S., Cao, Q., Aierken, A., Lee, H., and Lettenmaier, D. P. (2017), How much groundwater did California's Central Valley lose during the 20122016 drought?, Geophysical Research Letters, 44, 4872 4879, doi:10.1002/2017GL073333.

Wanders, N., A. Bachas, X.G. He, H. Huang, A. Koppa, Z.T. Mekonnen, B.R. Pagan, L.Q. Peng, N. Vergopolan, K.J. Wang, M. Xiao, S. Zhan, D.P. Lettenmaier, and E.F. Wood, 2017: Forecasting the Hydroclimatic Signature of the 2015/16 El Nio Event on the Western United States. Journal of Hydrometeorology, 18, 177186, https://doi.org/10.1175/JHM-D-16-0230.1

Becker, R., Koppa, A., aus der, B. T., Usman, M., Schulz, S., and Schuth, C. (Under Review). Automated model calibration of a highly managed hydrological system using remotely sensed evapotranspiration data. Journal of Hydrology

Koppa, A., Alam, S., Miralles, D., and Gebremichael, M. (Under Preparation). Global evaluation of precipitation and evapotranspiration datasets from a water and energy balance perspective.

CHAPTER 1

Introduction

1.1 Background and motivation

Hydropower is the largest source of renewable energy in the world, accounting for approximately 70 percent of the total power supply (Moran et al., 2018). However, the latter half of the 20^{th} century witnessed a steady decline in the number of operational hydropower dams in regions such as North America and Europe (O'Connor et al., 2015). In contrast, there has been a significant expansion of hydropower capacity, in the form of large and small dams, in developing regions such as India (Sharma et al., 2013), China (Chang et al., 2010), South East Asia (Stone, 2011), and Africa (Conway et al., 2015, 2017). The rapid increase in hydropower capacity in these regions has been characterized by low capacity utilization and inefficient operation of hydropower dams and cascades. One of the main reasons for the low productivity of hydropower dams in developing regions is the lack of accurate forecasts of inflow into reservoirs at various timescales. The inaccuracy of forecasts are a direct consequence of the severe scarcity of reliable hydrologic measurements in these regions (Figure 1.1). The reduction in the number of rainfall and streamflow gages (Stokstad, 1999) seen in recent years is expected to exacerbate the data-scarcity situation.

The primary objective of this dissertatation is to investigate the potential of satellite-based remote sensing in mitigating the issue of data-scarcity in studies concerning hydrologic modeling, forecasting, and hydropower optimization. Specifically, this dissertation aims to improve seasonal hydropower planning in data-scarce regions by leveraging global satellite-based remote sensing data products of hydrologic fluxes and storage variables. The dissertation is primarily motivated by the fact that most water balance components can be monitored

by satellite-based remote sensing, including precipitation, evapotranspiration, soil moisture, total water storage, and snow water equivalent (Lettenmaier et al., 2015). However, owing to differences sensor types and retrieval algorithms, a large number of global data products exists with varying accuracy in different topographies, geographies, and hydroclimates. For example, a global evaluation of nine satellite-based precipitation (Beck et al., 2017) and evapotranspiration (Miralles et al., 2016) datasets show a wide variance in the performance of different datasets. Therefore, the central motivational question that the dissertation tries to answer is: How can the wealth of satellite-based remote sensing datasets be utilized to improve seasonal hydropower planning in data-scarce regions? An overview of seasonal hydropower planning and its components is presented. In addition, an overview of challenges specific to data-scarce regions is enumerated. Then, the specific research questions addressed in this dissertation is detailed.

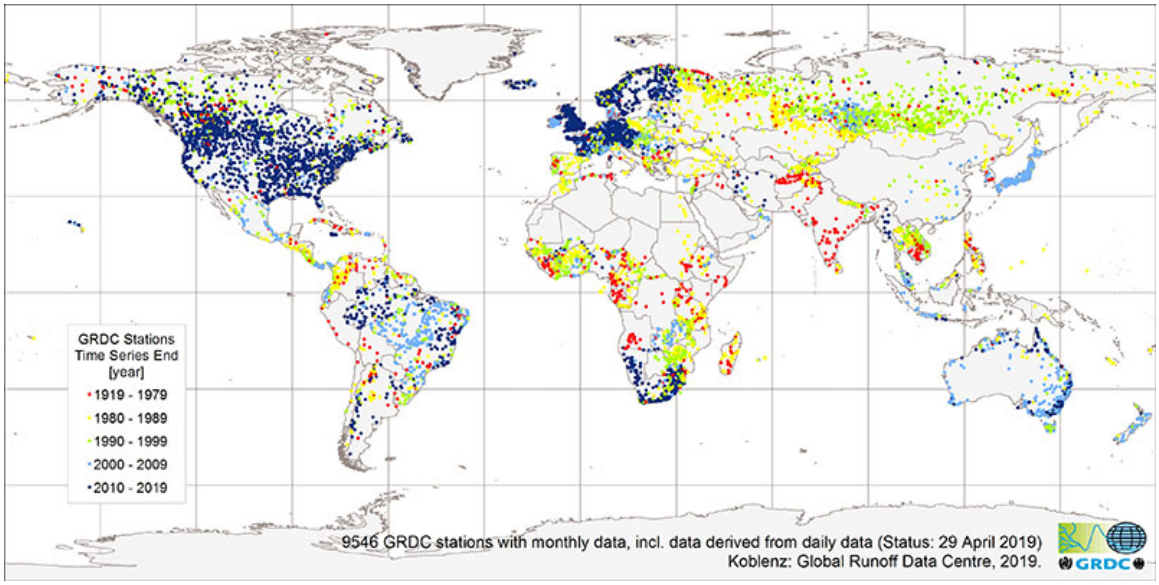

Figure 1.1: Spatial and temporal distribution of streamflow gage stations available in the Global Runoff Data Centre (sourced from https://www.bafg.de/).

1.2 Seasonal hydropower planning in data-scarce regions

Reservoir operations are managed at three distinct time-scales, serving different purposes: 1) long-term, 2) seasonal, and 3) real-time operations (Lund, 1996). Long-term studies are for developing general guidance and targets for the operation of the reservoir system. In the context of hydropower operations, it can be for deciding seasonal storage and release targets to achieve a certain hydropower target. These studies are carried out very infrequently. Within the context of long-term operations, seasonal operation planning is conducted monthly, semi-annually or annually to address shorter term conditions such as monthly release and storage decisions to fulfill the objectives and constraints of the reservoir or reservoir system. Real-term operation studies address issues such as short-term hydropower release, flood control, and water delivery requirements within thin the context of seasonal operations. This dissertation is focused on improving hydropower operations at seasonal time-scales using satellite-based estimates of hydrologic fluxes and storages.

A typical seasonal hydropower planning system consists of three main components or sub-systems linked together, with each component performing a distinct function: 1) climate system, 2) hydrology system, and 3) optimization system (Block, 2011). A visual representation of a seasonal hydropower planning system is presented in Figure 1.2. Seasonal forecasts of meteorological variables, typically precipitation and temperature, are translated into seasonal forecasts reservoir inflows using statistical or physically-based methods. The reservoir inflow forecasts acts as an input into a reservoir optimization model which generates optimized release decisions by maximizing a specified objective (for example, maximization of hydropower). Based on the type of inflow forecasts used, the seasonal hydropower planning can be deterministic or stochastic (Yeh, 1985).

In data-scarce regions, the lack of availability of reliable hydrologic information affects the functioning of each of the three components of seasonal hydropower planning. In the subsequent sections, an overview of each component is provided. More importantly, issues specific to data-scarce regions which hinder the generation of reliable precipitation forecasts, reservoir inflow forecasts, and optimization of hydropower at seasonal time-scales are iden-

tified. In addition, the role of satellite-based remote sensing in improving each of the three components of seasonal hydropower planning is discussed. Finally, the novel contributions of the dissertation towards addressing the identified issues is detailed.

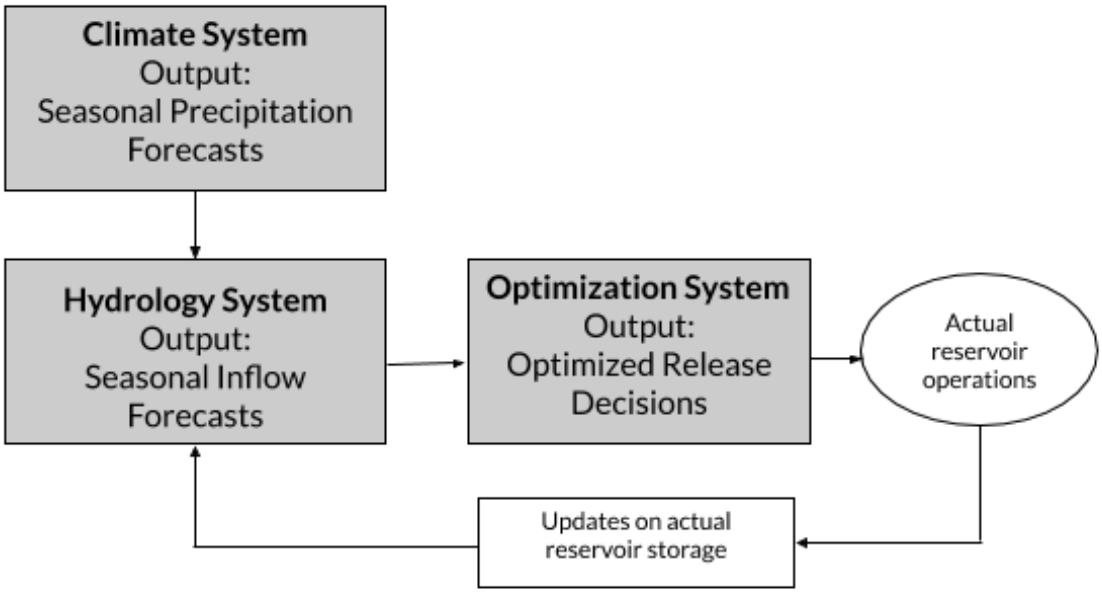

Figure 1.2: A representation of a seasonal hydropower planning system consisting of the climate, hydrology and optimization components and their linkage.

1.2.1 Climate system

The primary objective of the climate system is generate reliable seasonal forecast of precipitation. Seasonal forecasts of precipitation are generated by statistical and physcially-based approaches (Gerlitz et al., 2016). Statistical approaches try to exploit the relationship between precipitation and large scale climate anomalies such as the El Nino-Southern Oscillation (ENSO). This is especially true for tropical regions where ENSO governs the precipitation variability (Ratnam et al., 2014; Krishnaswamy et al., 2015). In the northern hydroclimatic regions such as Europe other climate anomalies such as the Indian Ocean Dipole (IOD) and Pacific Decadal Oscillation are found to be more influential in the context of seasonal precipitation forecasts (Eden et al., 2015). On the other hand physically-based forecasts

are produced by process-based dynamic models of atmosphere, termed as general circulation models (GCMs). These models solve fundamental equations which govern fluid dynamics to simulate the climate system. Typically run on a coarse grid, these models parameterize sub-grid scale processes such as convection. These models are used for seasonal foecasting by using observations of the current state of the atmosphere as initial and boundary conditions. Due to uncertainty in initial and boundary conditions, model parameter and physics, and the chaotic nature of atmsophere (Lorenz, 1963), the reliability of seasonal forecast from dynamic models is low (Kumar et al., 2013). To account for the aforementioned uncertainties, multi-model ensemble (MME) forecasting systems (refer to Chapter 4 for a detailed review of different multi-model ensembles). The dissertation is focused on multi-model ensemble seasonal forecasts of precipitation, rather than statistically generated foreacasts.

A number of postprocessing approaches are adopted improve the reliablity and accuracy of dynamical forecasts. The basic principle of postprocessing is to develop statistical relationships between forecasts and true climate represented by observations (Finnis et al., 2012). Postprocessing approaches include forecast calibration using linear (Jia et al., 2010), nonlinear regression (Zeng et al., 2011), and Bayesian methods (Herr and Krzysztofowicz, 2015). Postprocessing of multi-model ensemble forecasts include calibration and merging of ensemble members using skill-based weighting schemes (Schepen et al., 2016) and Bayesian model averaging (BMA) (Raftery et al., 2005). Irrespective of the postprocessing methodology adopted, the performance of these approaches depend on the quality of precipitation observations used a representation of the true climate. This plays a particularly important role in data-scarce regions where no reliable ground-based measurements of precipitation data are not available. In such regions, satellite remote sensing-based precipitation datasets can help address the data-scarcity issue. However, as hinted earlier, remotely sensed precipitation is subject to large temporal and spatial uncertainty. Therefore, the selection of reliable precipitation data is an important first step in improving the proliferation of remote sensing datasets for hydrologic and water resources studies in data-scarce regions. In Chapter 2, this dissertation attempts to address the issue of the selection of reliable remote sensing datasets

in data-scarce regions.

1.2.2 Hydrology system

The objective of the hydrology component of the seasonal hydropower planning system is to generate reliable reservoir inflow forecasts. Although several statistical approaches exist for seasonal streamflow (or inflow) forecasting (Piechota et al., 1998; Souza Filho and Lall, 2003; Gado Djibo et al., 2015), the dissertation focuses on dynamic methods utilizing physically-based rainfall-runoff models. The roots of physically-based seasonal streamflow forecasting can be traced to the development of the ensemble streamflow prediction (ESP) system by the National Weather Service in the United States (Day, 1985). In this method, a hydrologic model initialized with observed catchment conditions (such as snow water equivalent and soil moisture) is forced by an ensemble of precipitation scenarios sampled from historical data to produce probable realizations of future streamflow (or reservoir inflows). It is also possible to force the hydrologic model with NWP-based ensemble forecasts of precipitation. Irrespective of the forcing data, seasonal streamflow forecasting requires a robust model which can represent the hydrological characteristics of the study catchments with a reasonable degree of accuracy.

The setup of such a hydrologic model involves 1) accurately estimating model parameters (Shi et al., 2008) and 2) accurately representating catchment initial conditions (Koster et al., 2010). Chapter 4 presents a detailed discussion on sources of uncertainty in the hydrologic model and their effect on streamflow forecasts. Accurate estimation of model parameters (or calibration) requires reliable measurements of streamflow in the study region. In data-scarce catchments the calibration of such hydrologic models presents an unique challenge as streamflow measurements are not available and no reliable estimates of river flow are available from remote sensing-based sources (Lettenmaier et al., 2015). It is seen that calibration of hydrologic models can impact the accuracy of streamflow forecasts (Shi et al., 2008). Therefore, the dissertation identifies calibration of hydrologic models as a major challenge in generating reliable streamflow foreacsts in data-scarce regions. Chapter 3 presents a methodological

framework developed in this dissertation to identify reliable proxies for streamflow, such as evapotranspiration or soil moisture, which can be used to calibrate the hydrologic models.

1.2.3 Optimization system

The reservoir inflow forecasts generated by combining ensemble seasonal precipitation forecasts and rainfall-runoff model is used as an input to the optimization system to produce optimal release decisions to achieve a pre-determined objective (or objectives). The optimization of a reservoir or a system of reservoirs can be a deterministic or stochastic problem based on the nature of inflow forecasts used. As the dissertation uses an ensemble of seasonal precipitation forecasts to generate reservoir inflows, the focus is on stochastic optimization. In either case, optimization of reservoirs is formulated as a linear, nonlinear, or dynamic programming problem which consists of maximizing or minimizing a set of objective functions subject to a set of constraints (Labadie, 2004). Stochastic optimization is carried out under the assumption that the inflow forecasts are not perfect. A generalized objective function in stochastic optimization can be represented as

$$\max_{r} E_q[\sum_{t=1}^{T} \alpha_t f_t(s_t, r_t, q_t) + \alpha_{T+1}\phi_{T+1}(s_{T+1})] \tag{1.1}$$

where E = statistical expectation operator, r_t = n-dimensional set of control or decision variables during period t, T = length of the operational time horizon, s_t = n-dimensional state vector of storage in each reservoir at the beginning of period t, $f_t(s_t, r_t)$ = objective to be maximized or minimized, $\phi_{T+1}(s_{T+1})$ = final term representing future estimated benefits or costs beyond time horizon T, and α_t = discount factors for determining present values of future benefits or costs, q = inflows considered as random variables. Typically, the constraints are based on preserving the mass balance of the system of reservoirs, maintaining explicit lower and upper bounds of storage required for safe operation of the reservoirs. Additional constraints in the context of hydropower optimization is to maintain the releases within the capacity of turbines. The objective function and the constraints used in the dissertation for maximization of hydropower in the study region is presented in Chapter 5. The specific

issues involved in optimization of hydropower in data-scarce regions, the methodological framework developed in this dissertation to overcome these issues is presented in Chapter 5.

1.3 Organization of the dissertation

The dissertation is divided into five chapters. Chapter 1 provides an introduction discussing the background and motivation for the dissertation. Chapter 2 presents a novel validation framework for evaluating remotely sensed precipitation and evapotranspiration datasets without the need for ground-based measurements. Chapter 3 presents a framework for calibration of large scale hydrologic by combining Bayesian and Pareto-optimality principles. Chapter 4 uses the results from Chapter 2 and 3 to develop a seasonal hydropower planning framework for data-scarce regions using remote sensing data, multi-model ensemble precipitation forecasts and stochastic programming. Finally, Chapter 5 discusses the overarching conclusions drawn from the research carried out in the dissertation and proposes possible new lines of inquiry in this research area.

CHAPTER 2

A Framework for Validation of Remotely Sensed Precipitation and Evapotranspiration Based on the Budyko Hypothesis

2.1 Introduction

Advances in remote sensing have enabled continuous monitoring of water and energy fluxes at spatial and temporal scales appropriate for a wide range of hydrologic applications (Lettenmaier et al., 2015), especially in ungauged basins (Lakshmi, 2004). However, these estimates are subject to large uncertainties arising, primarily, from differences in retrieval algorithms and sensor types (Kidd and Huffman, 2011; Gebregiorgis and Hossain, 2014). For example, the Precipitation Intercomparison Project - 3 (PIP-3), which evaluated a number of satellite-based precipitation products concluded that, in general, Passive Microwave (PMW) based products are better than Infrared (IR) based products at shorter timescales (Adler et al., 2001). It is also seen that the performance of data products vary considerably in space, over different terrains. In a study by Hirpa et al. (2010), it was found that the Climate Prediction Center Morphing Technique (CMORPH) dataset performed satisfactorily over a complex terrain in Ethiopia whereas the same dataset failed to capture the temporal and spatial variation of rainfall over an urban region in China (Chen et al., 2014). Therefore, the use of remotely sensed data must be subject to comprehensive validation in the catchments of interest.

But, validation efforts are often hindered by the lack of ground-based observations in ungauged and data-scarce catchments. With substantial decline observed in in-situ pre-

cipitation and runoff measurement networks across the globe (Stokstad, 1999; Lorenz and Kunstmann, 2012), robust methodologies that can test the physical consistency of observational data using hydrologic principles can offer insight into their quality. Several studies make use of physically-based hydrologic models to validate remote sensing data. These studies are based on the hypothesis that any error in the input precipitation will result in a commensurate error in the closure of water balance. An effort to carry out the water balance of the Mississippi river basin using data exclusively from satellite remote sensing revealed that biases in precipitation resulted in overestimation of runoff (Sheffield et al., 2009). Stisen and Sandholt (2010) used a distributed hydrological model, calibrated using observed discharge, to evaluate precipitation datasets derived from satellite-borne sensors over a data-scarce catchment in Senegal. In addition to requiring streamflow observations, the main drawback of such validation studies is the use of calibrated and over-parameterized models which prohibit drawing meaningful conclusions regarding the dataset (Bitew and Gebremichael, 2011).

In this study we address the issues of requiring concurrent streamflow observations, calibration and over-parameterization of models by invoking the Budyko hypothesis (Budyko, 1974) for validating remotely sensed water and energy balance components, specifically precipitation and evapotranspiration. The Budyko hypothesis is a semi-empirical model that describes the long-term combined water and energy balance of catchments. In recent years, the hypothesis has witnessed widespread application in climate change (Greve et al., 2014), land use change (Zhou et al., 2015) and ecohydrological (Gentine et al., 2012) studies. Through this study we intend to rigorously test the applicability of the Budyko hypothesis to validation of observational data derived from remote sensing. Specifically, we intend to answer the following questions: 1) Can the Budyko hypothesis be used to develop a data validation framework that does not require concurrent ground-based measurements? 2) How does the validation framework compare with traditional methods of evaluation that make use of in-situ measurements?

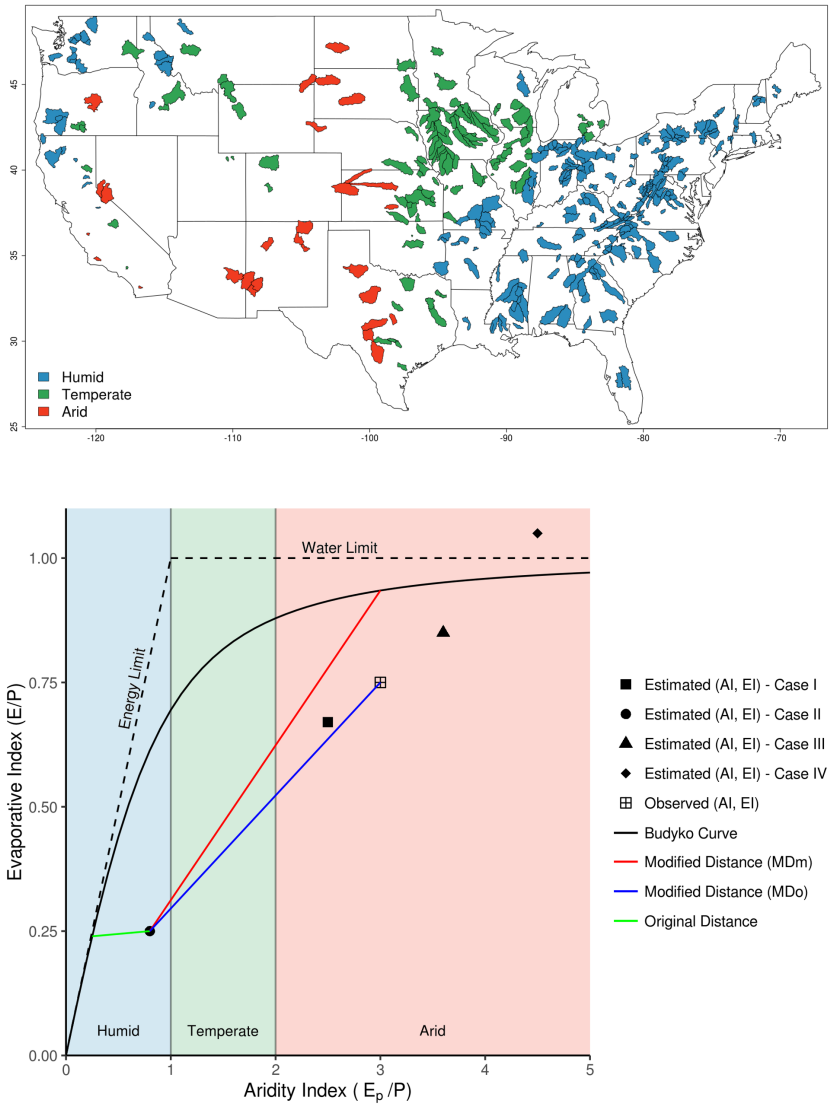

Figure 2.1: a) MOPEX catchments in the United States classified according to aridity and b) A representation of the Budyko space consisting of Fu's curve with $\omega = 2.6$ (Fu, 1981), water and energy limits. The original definition of distance following Greve et al. (2014) and the modified distances (MDm and MDo) are represented. Four cases of estimated (AI, EI) points representing different precipitation biases are shown: Case 1 - Low positive bias, Case 2 - High positive bias, Case 3 - Low negative bias and Case 4 - High negative bias (Refer to section 3.4).

2.2 Study area and observational data

For the development and validation of the proposed framework, the Model Parameter Estimation Experiment (MOPEX) catchments, consisting of 438 small to medium size basins spread across the United States are selected (Fig. 2.1a). For these catchments MOPEX provides ground-based daily observations of precipitation (P), runoff (Q) and climatological potential evapotranspiration (E_p) spanning 56 years (1948 to 2003) (downloaded from ftp://hydrology.nws.noaa.gov/pub/gcip/mopex/). The size of the catchments vary from 70 km^2 to 10,000 km^2. Based on continuous availability of data for at least 30 years, 393 out of 438 catchments were selected for the development of the framework. Evapotranspiration (E) is calculated as (P-Q) at annual time scales based on previous studies on MOPEX catchments (Li et al., 2013; Greve et al., 2015).

2.3 Development of the validation framework

2.3.1 Budyko Space

The Budyko space is a two-dimensional space in which every point is described by two dimensionless indices; Evaporative Index (E/P, abbreviated as EI) and Aridity Index (E_P/P, abbreviated as AI) (Fig.2.1). Physically, EI represents the partitioning of P into E in the long-term whereas AI can be interpreted as a measure of the mean climate of the catchment (Carmona et al., 2016). The original Budyko formulation (Budyko, 1974) relates AI to EI through a non-parametric, nonlinear function that describes the long-term water-energy balance of catchments. But, owing to observed deviations from the original Budyko curve (Donohue et al., 2012; Istanbulluoglu et al., 2012; Porporato et al., 2004), several parametric variations of the Budyko function have been formulated (Fu, 1981; Choudhury, 1999; Zhang et al., 2004; Yang et al., 2008). In this study, Fu's equation (Fu, 1981), a single parameter Budyko function is made use of (Fig. 2.1). It is expressed as follows:

$$\frac{E}{P} = 1 + \frac{E_p}{P} - \left(1 + \left(\frac{E_p}{P}\right)^{\omega}\right)^{\frac{1}{\omega}} \tag{2.1}$$

In equation (3.6), the parameter ω has no analytic solution but various parameterizations have been proposed that relate ω to catchment characteristics such as vegetation, soil, topography and location of catchments (Li et al., 2013; Xu et al., 2013).

The Budyko curve (Eq. 3.6) is constrained by the following water and energy limits:

$$\frac{E}{P} = 1, \quad \frac{E_p}{P} > 1 \quad (water\,limit) \tag{2.2}$$

$$\frac{E}{P} = \frac{Ep}{P}, \quad \frac{E_p}{P} < 1 \quad (energy\,limit) \tag{2.3}$$

Equation (2.2) implies that the water available for E is limited by P. Thus, the hypothesis assumes that the contribution of catchment storage to water availability is negligible over the long term. If the supply of water in the catchment is unlimited, E is constrained by available energy in the form of E_p (Eq. 2.3). In summary, the Budyko space consists of Budyko curve, based on catchment specific parameter ω, and the limits (Eq. 2.2 and 2.3).

2.3.2 Defining the error metric

In traditional data validation approaches, observational datasets are directly evaluated with the help of statistical error metrics such as Mean Bias and Root Mean Square Error (RMSE). Even when models are used to evaluate remote sensing datasets, the error in the dataset is assessed by comparing model outputs, like runoff, with ground-based measurements using similar error metrics. But using the Budyko function (Eq. 3.6) for validation requires the projection of the three-dimensional space described by (P, E, E_p) onto the two-dimensional Budyko space represented by (AI, EI). Therefore, the objective of this study is to develop and test an error metric that can translate the individual errors in P, E and E_p to an equivalent error in the Budyko space. For this, we use the Root Mean Square weighted Error (RMSwE) metric developed by Greve et al. (2014) with suitable modifications. This approach hypothesizes that the error in the (P, E, E_p) space is analogous to the euclidean distance between the estimated (AI, EI) point (Fig. 2.1b), determined from satellite-based estimates of (P, E and E_p), and the catchment specific Budyko curve (Original Distance in Fig.2.1b). Physically, Original Distance (OD), as defined by Greve et al. (2014), represents

the the combined error of precipitation and evapotranspiration datasets in representing the long-term water and energy balance of the catchments. In addition, a weight that penalizes the datasets that exceed the water and energy limits (Eq. 2.2 and 2.3) is introduced.

In this study, we set the value of the weight as unity based on the following arguments. One of the main objectives of this study is to compare the proposed framework with traditional data validation methods which do not contain any additional weight or penalty terms for exceeding water or energy limits. Therefore, comparing such an error metric with a metric that is inflated by an extraneous penalty would not be logical. The developed error metric may imply higher error in the P, E or E_p data than that would be implied by traditional error metrics which may lead to erroneous conclusions regarding the quality of the dataset. Hence, RMSE_{OD} in the Budyko space, following Greve et al. (2014) is expressed as

$$RMSE_{OD} = \sqrt{\frac{\sum_{i=1}^{n}(w_i.D_i)^2}{n}} \tag{2.4}$$

where the weight, $w_i = 1$, n is the number of points in the Budyko space and D_i or OD is the euclidean distance (minimum) between the point i and the catchment specific Budyko curve, RMSE_{OD} is the RMSE metric as defined by Greve et al. (2014) in which the subscript OD refers to the original definition of distance, D (refer to Fig. 2.1).

2.3.3 Estimating the error metric

Estimation of the RMSE_{OD} metric (Eq. 2.4) for the MOPEX catchments involves the calculation of the euclidean or the minimum distance, OD, between the estimated (AI, EI) point determined using remote sensing data (Fig. 2.1) and the Budyko curve defined by Fu's equation (Eq. 3.6). Thus, OD, according to Greve et al. (2014) is given by

$$\min_{AI} OD = \sqrt{(AI_{est} - AI)^2 + (EI_{est} - EI_{mod})^2} \tag{2.5}$$

where AI_{est} and EI_{est} are long-term average AI and EI estimated from remote sensing data, EI_{mod} is modeled EI from Fu's equation given by $EI_{mod} = 1 + AI - (1 + (AI)^{\omega})^{1/\omega})$.

The following steps are followed to estimate the RMSE_{OD} metric for the MOPEX catchments. First, the parameter ω in Fu's equation is estimated for each of the 393 MOPEX

catchments under consideration following the method prescribed by Li et al. (2013). In this method, ω is determined by minimizing the squared error between observed annual EI and EI inferred from Fu's equation (Eq. 3.6). Thus, the objective function is given by,

$$obj = \min \sum_i \{EI^i_{act} - (1 + AI^i_{act} - (1 + (AI^i_{act})^{\omega})^{1/\omega})\}^2 \tag{2.6}$$

where EI^i_{act} and AI^i_{act} are the average EI and AI estimated for the year i using observed data from MOPEX catchments.

Next, the original distance, OD, is determined for all the 393 catchments using equation (2.5). Finally, RMSE_{OD} is estimated using equation (2.4).

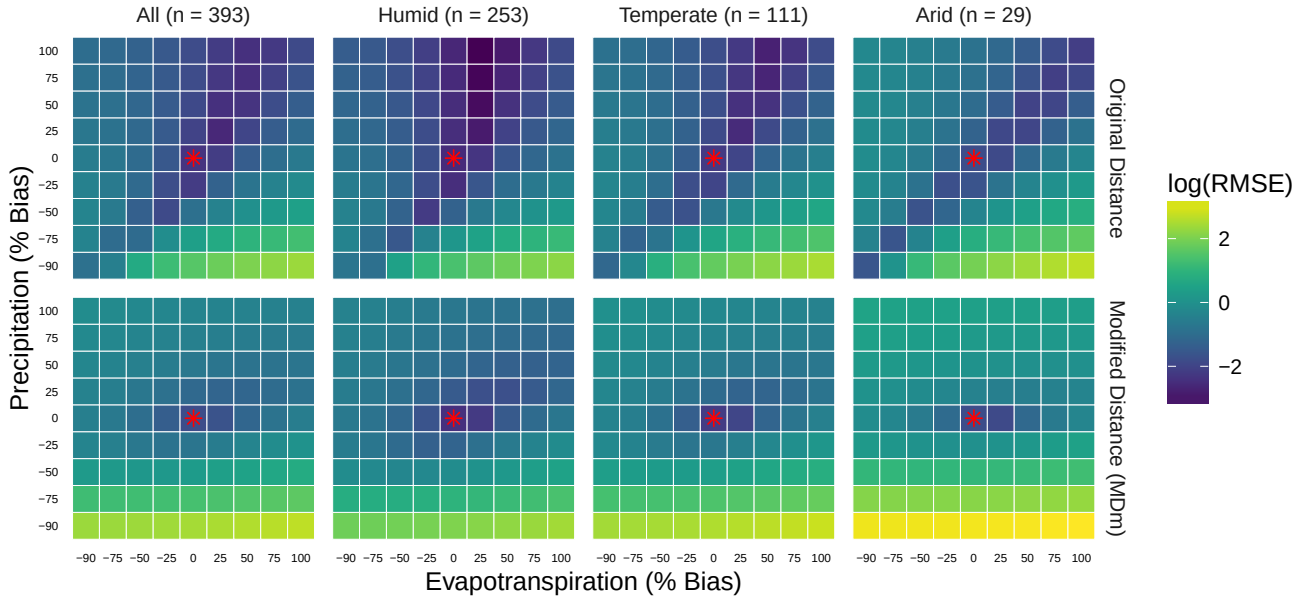

Figure 2.2: Sensitivity maps of the logarithm of 1) RMSE_{OD} (Original Distance) and 2) RMSE_{MDm} (Modified distance). The red star represents RMSE value for no bias in either precipitation or evapotranspiration. n is number of catchments in each hydroclimatic regime.

2.3.4 Sensitivity analysis of the error metric

We test the hypothesis that RMSE_{OD} in the Budyko space is sensitive to individual biases in the P and E estimates. For this, we setup a controlled sensitivity analysis experiment using

the 393 MOPEX catchments. We first divide the Budyko space into three hydroclimatic regimes based on AI; Humid ($0 < \text{AI} < 1$), Temperate ($1 < \text{AI} < 2$) and Arid ($\text{AI} > 2$) (Sankarasubramanian and Vogel, 2003). Out of 393 catchments, 253 are humid, 111 are temperate whereas 29 are arid (Fig. 2.1a). The observed P and E datasets provided by MOPEX are artificially biased by fixed percentages (0, 25, 50 etc) and for each combination of the biased P and E data, RMSE_{OD} (Eq. 2.4) is determined for the defined aridity classes. Then, the logarithm of the RMSE_{OD} value for each combination of artificially biased P and E data is plotted as a single pixel on the sensitivity map (Fig. 2.2). As the focus of the study is on validation of P and E datasets, the sensitivity of the error metric to changes in E_p was not considered

The sensitivity analysis results are interpreted using four cases - 1) Case I: Low positive precipitation bias, 2) Case II: High positive precipitation bias, 3) Case III: Low negative precipitation bias and 4) Case IV: High negative precipitation bias. These cases are represented in Figure 2.1. If the RMSE_{OD} metric is sensitive to individual biases in P and E values, then, ideally, the logarithm of RMSE_{OD} values must be minimum at the center of the sensitivity map (darker colors in Fig.2.2) and then gradually increase towards the edges (lighter colors in Fig. 2.2). It is evident from the sensitivity map of RMSE_{OD} (Fig. 2.2) that when the precipitation is positively biased to a large degree ($> 50\%$) the RMSE value is very low, irrespective of the hydroclimatic regime (Case II in Fig. 2.1) . This is contrary to expectation as RMSE value should increase with increasing bias in either P or E. The reason for such a behavior is that the euclidean distance, D, in equation (2.4) is the minimum distance between the Budyko curve and the estimated (AI,EI) point (Original Distance in Fig. 2.1b). As a result, if the precipitation has a large positive bias, the estimated (AI, EI) point moves towards the origin, or the humid region, of the Budyko space. In this region, distance to the Budyko curve is observed to be considerably lower than in either temperate or arid regions, thereby resulting in a very low RMSE_{OD} value. This is seen in cases I, III and IV as well in which the RMSE_{OD} metric behaves as expected when the estimated (AI, EI) point is not very close to the Budyko curve. Figure (2.2) also reveals that the RMSE_{OD}

metric is more sensitive to biases in P rather than E. This is expected as both AI and EI feature P in their definitions, whereas changes in E only affect the evaporative index (EI).

2.3.5 Modification of the error metric

We address the issue of high precipitation values leading to low RMSE by modifying the definition of the distance (D) in equation (2.4). Instead of determining the minimum distance from the estimated (AI, EI) point to the Budyko curve, D is calculated as the distance from the estimated (AI, EI) point to a point on the Budyko curve corresponding to the observed long-term aridity index of the specific catchment (Modeled Modified Distance in Fig. 2.1b). This modification preserves the aridity of the catchment which is violated by the original definition of distance but it introduces an additional data requirement in the form of the actual aridity index. The RMSE calculated using the modified definition of distance is henceforth referred to as RMSE_{MDm}, where MDm refers to modeled modified distance calculated as

$$MDm = \sqrt{(AI_{est} - AI_{act})^2 + (EI_{est} - EI_{mod})^2} \tag{2.7}$$

where EI_{est} and AI_{est} are the long-term average evaporative and aridity indices estimated from remote sensing data, AI_{act} is the actual AI determined using observed E_p and P data available for the MOPEX catchments (described in Section 2). EI_{mod} is modeled EI from Fu's equation (Equation 3.6) as $EI_{mod} = 1 + AI_{act} - (1 + (AI_{act})^{\omega})^{1/\omega}$.

Sensitivity analysis is performed on the RMSE_{MDm} metric to test whether the new definition can overcome the shortcomings of RMSE_{OD}. In the sensitivity maps of RMSE_{MDm} (Fig. 2.2), very low values are found only around the pixel representing RMSE_{MDm} for zero bias in P and E. The positive and negative bias regions in the heat maps exhibit increasing gradients from the center (point of no bias) towards the edges (points of maximum bias). This behavior of RMSE_{MDm} is essential if any error metric in the Budyko space is to reflect the errors in the individual estimates of P and E. Comparing the sensitivity maps of RMSE_{MDm} with that of the RMSE_{OD}, it is quite apparent that the magnitude of error is higher in all hydroclimatic regimes . As expected RMSE_{MDm} is still more sensitive to biases

in P rather than E. Across hydroclimatic regimes, RMSE_{MDm} values in the arid region is higher than in temperate or humid regions for the same percentage of bias. The biggest improvement is seen in the humid regions, where insensitive regions in the RMSE_{OD} heat map now exhibit higher sensitivity.

2.3.6 Sensitivity of modified distance to ω and AI

Determining the modeled modified distance (MDm in equation 2.7), and thus the application of the developed RMSE_{MDm} error metric, requires estimates of both the Budyko parameter, ω, and the long-term aridity index, AI, of the catchment. For data-scarce catchments where no ground-based observations of P and E are available, estimates of ω and AI (AI_{act} in equation 2.7) have to be obtained from other sources. As stated earlier, several parameterizations are available for ω (Li et al., 2013; Xu et al., 2013) and for AI_{act}, long-term aridity index maps are available (Trabucco and Zomer, 2009). As these parameterizations and estimates are subject to uncertainty, it is important to understand the importance of accurately estimating ω and AI_{act} to the determination of the developed RMSE_{MDm} metric. In this section, we answer the question: How sensitive is the modeled modified distance, MDm, in equation 2.7 to changes in 1) ω and 2) AI_{act}?

To answer the question, a theoretical experiment is setup in the Budyko space. First, three estimated (AI, EI) points (AI_{est} and EI_{est} in equation 2.7) representing the remotely sensed data are selected, one each in the humid ($\text{AI}_{est} = 0.5$, $\text{EI}_{est} = 0{,}5$), temperate ($\text{AI}_{est} = 1.5$, $\text{EI}_{est} = 0.5$) and arid regions ($\text{AI}_{est} = 4$, $\text{EI}_{est} = 0.5$) of the Budyko space. To quantify the effects of changes in ω, the value of the actual AI_{act} is fixed at 0.5, 1.5 and 4 for humid, temperate and arid regions of the Budyko space respectively. To quantify the sensitivity of the modified distance (MDm) to changes in AI_{act}, the value of ω is fixed at 2.6. Then, the sensitivity analysis is carried out according to Saltelli et al. (2008). Specifically, we adopt the stratified single parameter sampling strategy suggested by the authors. Accordingly, ω is sampled at equal intervals in the range [1,6]. The range of ω is selected based on the estimated ω values for the MOPEX catchments. AI_{act} is sampled in a similar manner in the

ranges [0,1] for humid, [1,2] for temperate and [2,6] for arid regions in the Budyko space. 2000 samples of ω and AI_{act} are obtained to estimate the mean and variance of the resulting distribution of modified distance, with variance being a measure of sensitivity.

Table 2.1 summarizes the results of sensitivity analysis. It is seen that irrespective of the position of the estimated (AI_{est}, EI_{est}) point or the aridity region in which sensitivity analysis is carried out, the variance of modified distance with varying AI_{act} is orders of magnitude higher than the variance in modified distance with varying ω. Therefore, modified distance is more sensitive to changes in AI rather than ω. In absolute terms, changes in the Budyko parameter does not seem to significantly affect the estimates of modified distance. Even for AI_{act}, the magnitude of the variance of the modified distance is higher in arid region compared to the other regions. This is due to the fact that the range of aridity index in the arid region is higher than the other regions and the non-linear nature of the Budyko curve. Therefore, when applying the framework to data-scarce regions it is important to have more reliable estimates of AI, especially if the catchment is in the arid region. Moreover, as the variance magnitudes are not very high, the uncertainty in RMSE_{MDm} estimates are relatively low (except for arid catchments). It is to be noted that sensitivity analysis was carried out for a number of different estimated (AI_{est}, EI_{est}) points in the Budyko space but the results are similar to the results presented here.

2.3.7 Statistical significance test

The difference in RMSE_{MDm} values between every combination of P and E datasets is tested for statistical significance using a two-sample Student's t-test (two-sided) (Snedecor and Cochran, 1989). To apply the t-test a sample of 100 RMSE_{MD} values are generated for each combination of P and E datasets. The large sample size, $N = 100$, is to overcome the requirement of normality for applying t-tests. Each of the 100 RMSE_{MDm} values are calculated by randomly selecting 150 catchments from the 393 MOPEX catchments. The null hypothesis in this case is $H0 : \mu_i = \mu_j$, where μ_i and μ_j are the means of RMSE_{MDm} values of two different combinations of P and E datasets. For example, μ_1 could be mean of

Table 2.1: Variance and mean (in brackets) of the modified distance (MDm) for 1) Fixed AI_{act} and varying ω and 2) Fixed ω and varying AI_{act}. Depending on the values of Estimated (AI_{est}, EI_{est}) and actual AI_{act}, 18 cases are presented. The fixed AI_{act} values for humid, temperate and arid regions are 0.5, 1.5 and 4 respectively. The fixed ω value is 2.6

I) Estimated (AI_{est} = 0.5, EI_{est} = 0.5) point in Humid region			
Aridity	**Humid**	**Temperate**	**Arid**
Fixed AI_{act} and varying ω	0.012 (0.09)	1.0E-03 (1.06)	1.2E-04 (3.52)
Fixed ω and varying AI_{act}	0.03 (0.32)	0.085 (1.05)	1.32 (3.53)
II) Estimated (AI_{est} = 1.5, EI_{est} = 0.5) point in Temperate region			
Aridity	**Humid**	**Temperate**	**Arid**
Fixed AI_{act} and varying ω	4.3E-04 (1.01)	0.02 (0.33)	2.4E-04 (2.54)
Fixed ω and varying AI_{act}	0.09 (1.02)	0.08 (0.42)	1.28 (2.55)
III) Estimated (AI_{est} = 4, EI_{est} = 0.5) point in Arid region			
Aridity	**Humid**	**Temperate**	**Arid**
Fixed AI_{act} and varying ω	3.8E-05 (3.50)	1.9E-04 (2.52)	0.015 (0.41)
Fixed ω and varying AI_{act}	0.09 (3.51)	0.08 (2.52)	0.24 (1.14)

$RMSE_{MDm}$ values of (TRMM, AVHRR) and μ_2 could be mean for (TRMM, MODIS). The threshold selected for statistical significance is $\alpha = 0.05$.

2.4 Validating the framework

For validating the framework, several satellite-based P and E datasets are used. For P estimates three datasets are chosen; Climate Prediction Center Morphing Technique (CMORPH) (Joyce et al., 2004) (downloaded from http://rda.ucar.edu/datasets/ds502.0), Tropical Rainfall Measuring Mission (TRMM) 3B42RT (Huffman et al., 2007) and the Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks - Climate Data Record (PERSIANN) (Ashouri et al., 2015). Three evapotranspiration products are used; Advanced Very High Resolution Radiometer (AVHRR) (Zhang et al., 2010) (from

http://www.ntsg.umt.edu/project/et), MOD16 from the Moderate Resolution Imaging Spectrometer (MODIS) (Mu et al., 2007) (from http://www.ntsg.umt.edu/project/mod16) and the Global Land Evaporation Amsterdam Model (GLEAM) (Miralles et al., 2011; Martens et al., 2016) (from http://www.gleam.eu/). As the study focuses on validation of P and E dataset, E_p is sourced only from GLEAM. As most of the remote sensing data span the years 1998-2015, the MOPEX dataset was also extended from 2003 to 2015 using precipitation from Parameter-elevation Relationships on Independent Slopes (PRISM) (PRISM Climate Group, Oregon State University, http://prism.oregonstate.edu) and river discharge data from the United States Geological Survey.

2.4.1 Validation in the Budyko space

To validate the framework, RMSE metric in the Budyko space modeled using Fu's equation (RMSE_{MDm}) is compared against RMSE metric in the Budyko space calculated using observed data (RMSE_{MDo}), to see whether both the metrics lead to similar conclusions regarding the quality of datasets under consideration. RMSE_{MDo} is calculated using equation 2.4, but the distance, D, is the observed modified distance (MDo) between the estimated (AI_{est}, EI_{est}) point and the actual (AI_{act}, EI_{act}) point (Fig. 2.1) given by

$$MDo = \sqrt{(AI_{est} - AI_{act})^2 + (EI_{est} - EI_{act})^2} \tag{2.8}$$

To determine the estimated (AI_{est}, EI_{est}) point, the long-term annual average estimates of P and E are determined from the aforementioned satellite-based products. To determine RMSE_{MDm}, the actual AI (AI_{act} in equation 2.7) is calculated using E_p from the MOPEX dataset and P from the MOPEX dataset. To determine RMSE_{MDo}, the actual (AI_{act}, EI_{act}) point (equation 2.8) is calculated using the extended MOPEX dataset.

It is evident that the developed error metric using Fu's equation (RMSE_{MDm}) compares well with the observed RMSE_{MDo} metric in characterizing the combined error in P and E datasets (Fig. 2.3) in all hydroclimates. The agreement between RMSE_{MDm} and RMSE_{MDo} is quite apparent in the scatter plot of observed and modeled distances (Fig. 2.4). Except for

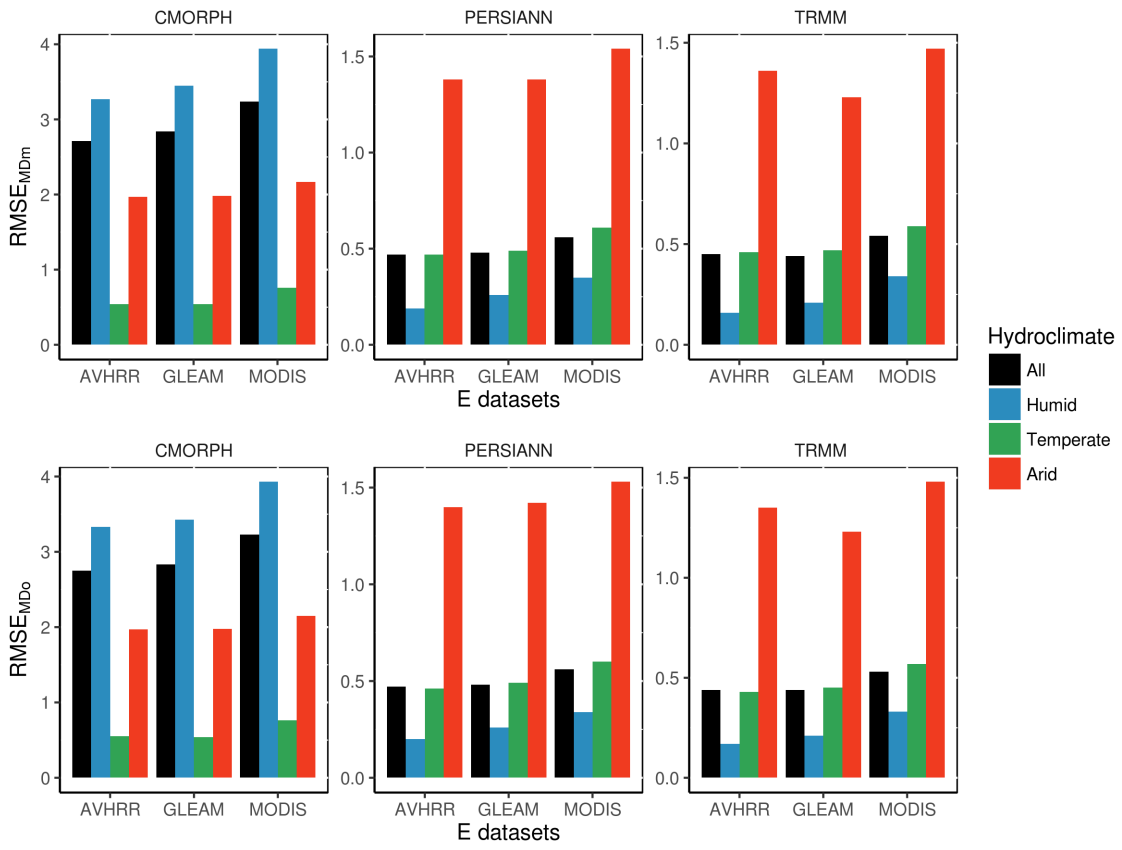

Figure 2.3: Bar Plots of RMSE_{MDm} and RMSE_{MDo} values for different combinations of P and E datasets and for all hydroclimatic regimes. The three panels correspond to the three satellite-based Precipitation products.

a few catchments in the arid and humid regions, the distances are well correlated. Analyzing RMSE_{MDo} values determined using ground-based measurements of P and E (Fig. 2.3) for all the hydroclimatic regimes combined, it is seen that the (P, E) combination that exhibits the least error is (TRMM, AVHRR). A similar conclusion is reached by the proposed Budyko hypothesis-based validation framework (Fig. 2.3). As seen earlier, RMSE_{MDm} exhibits different sensitivity in different hydroclimates. To test whether the conclusion regarding the best combination of (P, E) dataset holds true for different hydroclimates, we determine the RMSE_{MDm} and RMSE_{MDo} statistics separately for humid, temperate and arid catchments. The results show that both metrics point to (TRMM, AVHRR) as the best combination across all hydroclimates. As far as the poorest performing (P, E) datasets are concerned,

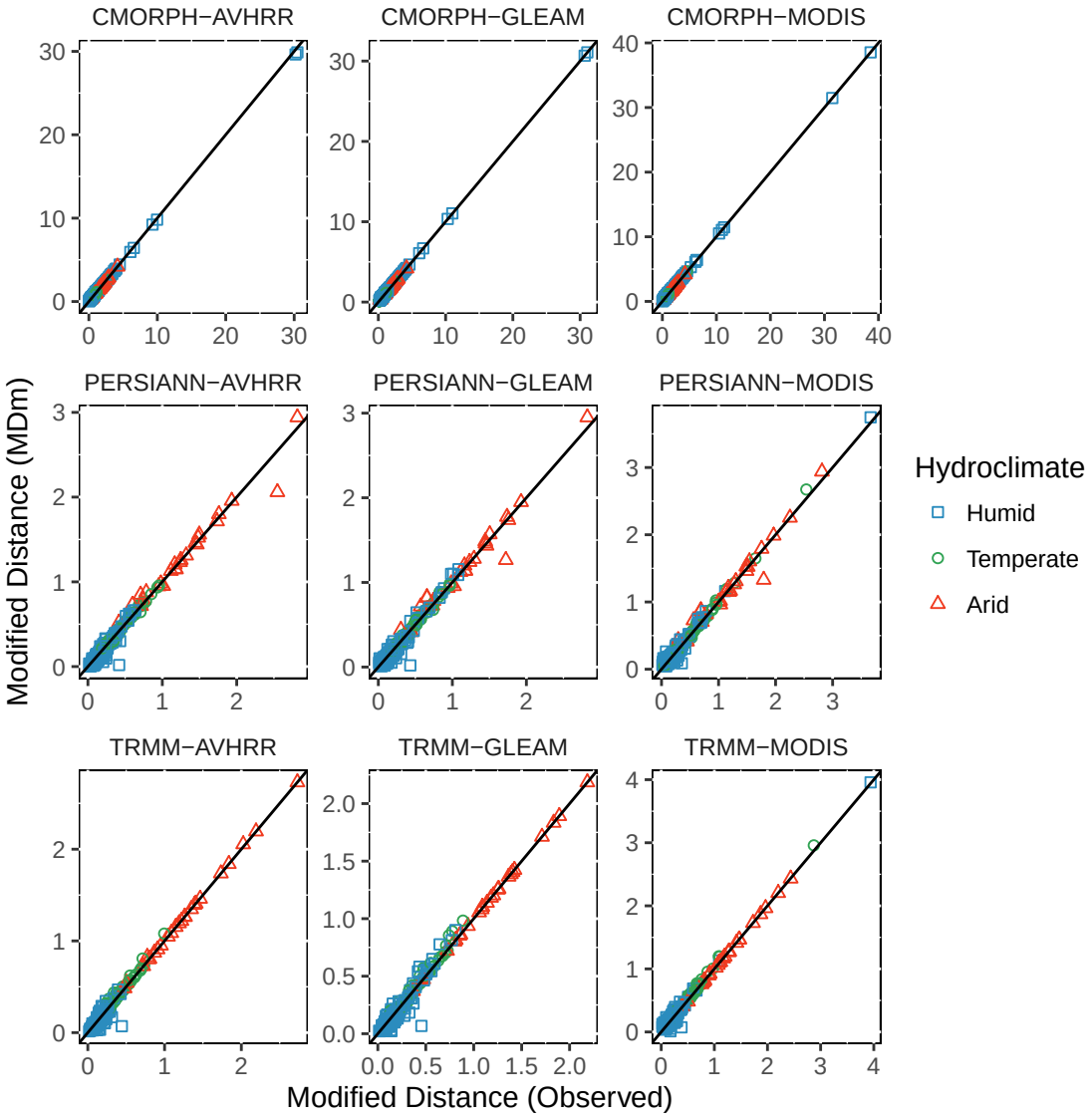

Figure 2.4: Scatter plots of observed modified distance (X-axis) versus modeled modified distance (Y-axis) for the MOPEX catchments and for all combinations of remotely sensed P and E datasets.

RMSE_{MDm} and RMSE_{MDo} metrics agree that (CMORPH, MODIS) combination fails to represent the combined water and energy balance of the study catchments. CMORPH seems to have high errors, especially in humid catchments (Fig. 2.3). On closer examination of the distribution of (AI, EI) points in the Budyko space for CMORPH, it is seen that in a number of catchments, the water and energy limits are exceeded by a large margin, mainly

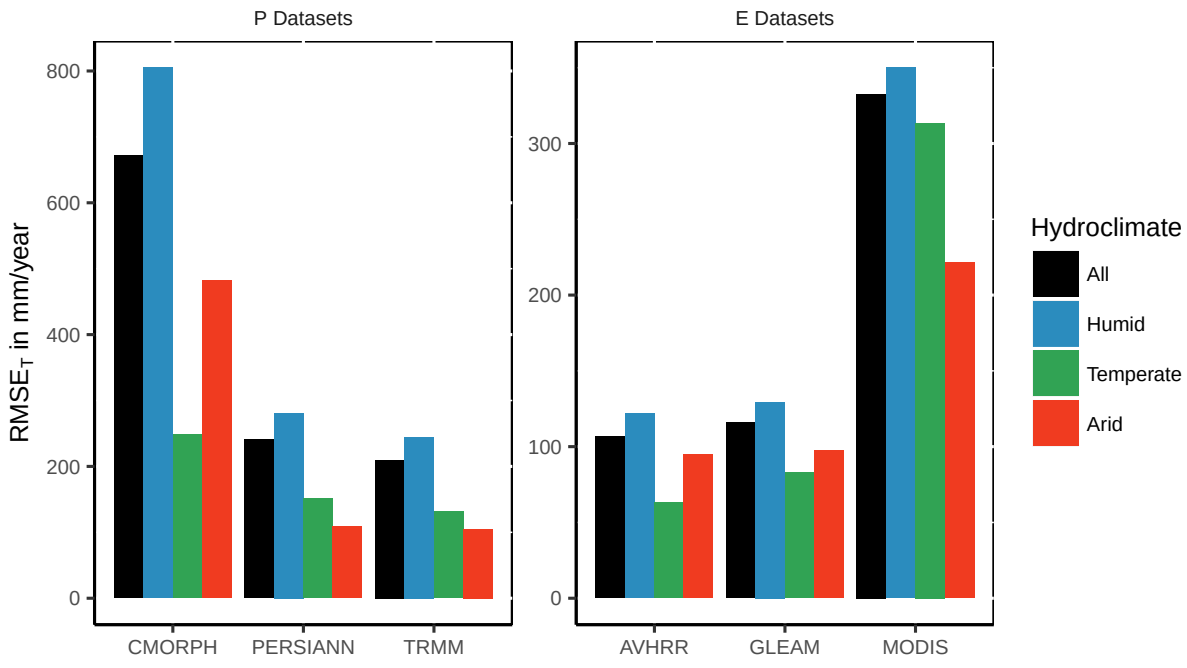

Figure 2.5: Bar Plots of RMSE$_T$ (Traditional RMSE) values for different P and E datasets and for all hydroclimatic regimes.

due to severe underestimation of precipitation. MODIS combined with P datasets results in higher RMSE$_{MDm}$ values compared to either GLEAM or AVHRR.

2.4.2 Validation with traditional RMSE

In general, the proposed error metric is seen to be capable of identifying the best performing combination of P and E datasets in the Budyko space. But it is also important to understand how the developed RMSE$_{MDm}$ metric compares with traditional RMSE$_T$ metric for several reasons. First, it is essential to analyze whether combining P and E datasets, as is done in the Budyko space, substantially differs from evaluating P and E separately, as is done traditionally. Secondly, RMSE$_{MDm}$ is seen to be less sensitive to changes in E compared to changes in P. To understand the effect of combining P and E datasets and the reduced sensitivity of the error metric to E, we compare RMSE$_{MDm}$ to traditional RMSE$_T$ values

(Fig.2.5). The long-term average from the extended MOPEX dataset and the satellite based estimates of P and E are used to determine RMSE_T of individual datasets across all the hydroclimatic regimes.

As far as the best (TRMM and AVHRR) and the worst (CMORPH and MODIS) performing P and E datasets are concerned, the RMSE_T metric agrees with the developed framework. Therefore, combining P and E datasets in the Budyko space does not seem to have substantial effect on ranking the P and E datasets. But, it is seen that RMSE_{MDm} does indeed suppress the biases in E datasets. For example, MODIS is seen to have a very high RMSE_T error in the arid region, but combining MODIS and CMORPH in the Budyko space, the error is comparable to other combinations in magnitude. Next, disaggregating the catchments according to AI reveals certain differences between RMSE_{MD} and RMSE_T metrics (Fig. 2.3 and 2.5). For example, if RMSE_{MDm} is followed, the (TRMM, AVHRR) combination performs best in humid catchments. But RMSE_T for TRMM is minimum in arid catchments and AVHRR is seen to be better in temperate regions than in either humid or arid catchments. This behavior is primarily due to the fact that the distance, D, in humid regions are relatively small for the same magnitude of error in (P, E, E_p) compared to distance in arid and temperate regions (Fig. 2.2). Therefore, care must be taken in comparing RMSE_{MDm} values calculated in different regions of the Budyko space.

2.4.3 Statistical significance of the results

Owing to a large combination of P and E datasets, only a summary of the results of the statistical significance tests is presented here. Except a few exceptions, the difference in RMSE_{MDm} values between all combinations of P and E datasets are seen to be significant with a p-value less than 0.05. The exception involves the AVHRR and GLEAM evapotranspiration datasets where the difference in RMSE_{MDm} values is statistically insignificant when AVHRR and GLEAM are combined with the same P dataset (For example, p-value is greater than 0.05 between TRMM-AVHRR and TRMM-GLEAM). This is also seen in figure (2.5) where the traditional RMSE_T values are close to each other. The results of

statistical significance tests show that despite being relatively more insensitive to biases in E, the RMSE_{MDm} metric can represent these biases given that their magnitude is high. For example, the difference between TRMM-GLEAM and TRMM-MODIS combinations is significant.

2.5 Applying the framework to a data-scarce catchment

We demonstrate the application of the proposed framework to a catchment in which no ground-based measurement of P and E are available. The steps involved in determination of the Budyko parameter (ω), the modified distance, D and the RMSE_{MDm} error metric using publicly available data are detailed.

The study area selected for applying the framework is the Omo Gibe river basin in East Africa. As all the satellite-based P and E products are gridded, we divide the catchment into regular grids of resolution 0.25^o x 0.25^o. Thus the Omo Gibe basin, with an area of 80,000 km^2, is divided into 592 grids and all the six satellite-based P and E products are interpolated onto the grids using nearest-neighbor interpolation.

Next step is to determine the Budyko parameter, ω, for each of the 592 grids. For this we use the parameterization developed by Xu et al. (2013). Specifically, we make use of the following multiple linear regression (MLR) model developed for small catchments.

$$\omega = 5.05722 - 0.09322lat + 0.13085CTI + 1.31697NDVI + 0.00003A - 0.00018elev \quad (2.9)$$

where lat is the latitude of each of the grid points in degrees, A is grid resolution in km^2, CTI is compound topographic index (Beven and Kirkby, 1979), $NDVI$ is the Normalized Difference Vegetation Index and $elev$ is elevation of the grids in meters. In equation (2.9) the topographic variables (lat, $elev$ and CTI) were determined using the Hydro-1K dataset (Lehner et al., 2006) (downloaded from http://hydrosheds.cr.usgs.gov/dataavail.php), $NDVI$ is derived from the Global Inventory Modeling and Mapping Studies (GIMMS) AVHRR NDVI3g dataset (https://nex.nasa.gov/nex/projects/1349/).

Determination of RMSE_{MD} requires the knowledge of long-term AI of the catchments.

It is to be noted that only a climatological estimate of AI of the study catchment is required and not concurrent observations corresponding to the study period. AI of the 592 grids are derived from the Global Aridity Index dataset (spatial resolution of 1km) developed by Trabucco and Zomer (2009) (from http://www.cgiar-csi.org/data/global-aridity-and-pet-database). Next the distance, D, in equation (2.4) is determined. For this the satellite-based estimates of (P, E, E_p) are mapped onto the Budyko space. Then, the distance, MDm is calculated using equation (2.7) and $\text{RMSE}_M Dm$ metric is determined using equation (2.4).

The RMSE_{MDm} metric is determined for all combinations of P and E datasets (Eq. 2.4): 0.135 (CMORPH, AVHRR), 0.149 (CMORPH, GLEAM), 0.876 (CMORPH, MODIS), 0.132 (PERSIANN, AVHRR), 0.147 (PERSIANN, GLEAM), 0.53 (PERSIANN, MODIS), 0.133 (TRMM, AVHRR), 0.139 (TRMM, GLEAM) and 0.462 (TRMM, MODIS). It is apparent that the MODIS E dataset performs the poorest among the E datasets. Unlike in the MOPEX catchments, all the RMSE_{MDm} values of all the three P datasets under consideration are close to each other, especially when combined with AVHRR and GLEAM E estimates. In summary, the best precipitation products according to the developed framework are TRMM and PERSIANN (PERSIANN-CDR). As far as the best E dataset is concerned AVHRR is seen to have the least bias. The combination that shows the largest deviation from the Budyko curve is (CMORPH, MODIS). The results from the developed framework are compared with past studies which validate satellite products for Ethiopian river basins. Romilly and Gebremichael (2011) and Hirpa et al. (2010) concluded that TRMM 3B42RT outperforms CMORPH in Ethiopian basins. As both the aforementioned studies evaluate the PERSIANN-CCS product and not the PERSIANN-CDR multi-satellite product used in the present study, the result from the developed framework for PERSIANN could not be validated with past studies.

It is to be noted here, that the methodology presented here is just one of the ways in which this framework can be applied. This application assumes that no ground-based measurements of P and E for the study period is available. But in data-scarce catchments, where some historical data is available, the Budyko parameter ω and aridity index, AI, can

be estimated using the methods presented in the paper for MOPEX catchments.

2.6 Conclusions, limitations and future work

Hydrologic studies in data-scarce regions use remotely sensed precipitation as meteorological forcing and satellite-based estimates of evapotranspiration for data assimilation. But remotely sensed P and E datasets exhibit large uncertainty requiring comprehensive validation for the area of study. This study addresses this important issue by developing a validation framework that tests for the physical consistency of remotely sensed P and E datasets without the use of concurrent ground-based measurements. A RMSE-based error metric is developed and comprehensively tested to see whether the metric can translate individual biases in P and E datasets onto the Budyko space. Results show that the proposed validation framework is capable of arriving at the same conclusions as traditional validation methodologies regarding the quality of P and E datasets. The application of the developed framework to a data-scarce catchment using publicly available topographic, vegetation and aridity information is also presented. In contrast to previous validation studies that employ complex distributed hydrological models, the use of the single parameter Budyko function highlights the effectiveness of using simple water and energy balance principles in validation of observational data.

Owing to the limitations of the original Budyko formulation, the developed framework can only test whether the combination of P and E datasets can describe the long-term combined water and energy balance of catchments. This implies that the developed RMSE_{MDm} metric characterizes the bias in P and E datasets and not the variance. Recent studies have focused on extending the Budyko hypothesis to sub-annual timescales (Greve et al., 2016; Zhang et al., 2008). Therefore, future work involves the extension of the framework to validation of P and E datasets at monthly and daily time-scales, which is crucial for characterizing the variance and also for hydrologic applications such as streamflow forecasting and reservoir operations. In addition, it is assumed that in the long-term, storage in the catchment is negligible. Therefore, care must be taken when applying the framework in catchments which

have long-term storage such as snow, ice or reservoirs and also in small catchments where storage influences water availability. It is seen that the RMSE_{MDm} metric is more sensitive to biases in P rather than E. Therefore, care must be taken in interpreting the error metric when the focus of a study is solely on evaluating E datasets which are relatively close to each other. But E datasets can be still be effectively evaluated using the framework if accurate estimates of P are available and the focus of the study is to validate only E datasets

It is to be noted here that the developed framework does not require concurrent observations of precipitation and evapotranspiration for validating remote sensing data. But the application of this framework to a data-scarce region requires reliable estimates of AI, which could be sourced from non-concurrent ground-based measurements, as was done in this study. Although a large sample of catchments, representing a wide range of aridities, have been used in the study, we encourage researchers to validate the robustness of the developed framework in other geographies having different topographic, hydrologic and climatic characteristics.

CHAPTER 3

Calibration of Large Scale Hydrologic Models with Multiple Fluxes: The Necessity and Value of a Pareto Optimal Approach

3.1 Introduction

The widespread use of large scale hydrologic and land surface models (LSMs) in snow (Christensen and Lettenmaier, 2007; Li et al., 2017), drought (Leng et al., 2015; Sheffield et al., 2004), and climate change (Middelkoop et al., 2001; Cuo et al., 2013) studies necessitates critical examination of the adopted calibration methodologies. The general approach of calibrating the models with measurements of a single flux, typically streamflow, is considered inadequate for such studies that require other water balance components to be simulated accurately. Rakovec et al. (2016b) evaluate the performance of the mesoscale hydrologic model (mHM) calibrated with streamflow (SF) against observed evapotranspiration (ET), soil moisture (SM), and total water storage (TWS). The study concludes that calibrating hydrologic models with only streamflow may not be sufficient for accurate simulation of other water balance components. Wanders et al. (2014) calibrate the LISFLOOD hydrologic model with remotely sensed soil moisture datasets. The results of the study show that calibrating the hydrologic model with only soil moisture negatively affects the accuracy of the corresponding streamflow simulation (compared to streamflow-calibrated model results). López López et al. (2017) confirm the findings of the other studies; calibrating a hydrologic model with only ET or SM adversely affects the accuracy of streamflow simulation compared to streamflow-calibrated model results. In Zink et al. (2018), a land surface model calibrated

with land surface temperature leads to higher errors in streamflow simulations compared to a streamflow-calibrated model.

Incorporating multiple fluxes into the calibration process has emerged as a consensus solution to address the adverse effects of single objective calibration. A number of different methods have been employed to calibrate hydrologic and land surface models with multiple fluxes, including stepwise calibration (Sutanudjaja et al., 2013; López López et al., 2017), ensemble Kalman filter (Wanders et al., 2014), and simultaneous calibration by combining objective functions (Rientjes et al., 2013; Rakovec et al., 2016a; Zink et al., 2018). Irrespective of the calibration strategy adopted, all the studies report improvements in the simulation of the added flux or storage component while maintaining the accuracy of the primary variable of interest. The improvements are also consistent across different water balance components incorporated into calibration, including evapotranspiration (Rientjes et al., 2013; López López et al., 2017; Zink et al., 2018), soil moisture (Sutanudjaja et al., 2013; Wanders et al., 2014), and total water storage (Rakovec et al., 2016a). Although the enumerated studies provide evidence in favor of multivariate calibration, we identify shortcomings in these approaches that hinder comprehensive quantification of the value of incorporating additional fluxes.

First, multivariate calibration studies do not define any limits of acceptability or error thresholds to determine whether the model can simultaneously reproduce the incorporated fluxes to a sufficient degree of accuracy. To illustrate the importance of defining limits of acceptability, consider the results of Rakovec et al. (2016a) wherein the addition of TWS estimates into calibration along with streamflow reduce the root mean square error (RMSE) of TWS simulations at negligible cost to the accuracy of streamflow simulation. A closer analysis of the results reveals that despite reduction in the RMSE of TWS, the absolute value of RMSE is still significantly large (Figure 3. in Rakovec et al. (2016a)); whereas the RMSE of standardized anomalies of streamflow has a median of approximately 0.5, the RMSE of TWS is approximately 0.8 (reduced from 0.9 for the SF-calibrated model). Without defining a threshold for acceptable error, it is difficult to assess whether the reported reduction in

TWS error is sufficient evidence to conclude that the incorporation of an additional flux actually improves the realism of the model.

Second, most studies consider the relationship between different fluxes to be complementary but all the results, with the exception of Wanders et al. (2014), point towards a trade-off relationship. By definition, a complementary relationship would mean that incorporation of additional fluxes improves the accuracy of all the incorporated fluxes. The objective functions of calibration and the calibration methodologies (such as stepwise calibration) are constructed to reflect the assumption of a complementary relationship. Even in Wanders et al. (2014) the improvement in SF accuracy when SM is incorporated, compared with a streamflow-calibrated model, is limited to small catchments. In addition, the results of multivariate calibration rarely are compared with results of models calibrated only with the additional flux. Such a comparison would help in quantifying the potential trade-offs in simulating the two fluxes accurately. For example, in Rakovec et al. (2016a) the model is not calibrated with only TWS, which would help understand the trade-off in TWS accuracy required to achieve acceptable SF accuracy. In studies where all the calibration cases are reported, there are significant trade-offs among the different fluxes considered for calibration (Rientjes et al., 2013). Even in Zink et al. (2018), where the objective function is designed to produce a compromise solution between the SF and ET fluxes, there is no discussion of either the magnitude of trade-off in the accuracy of ET or of whether such trade-offs are within acceptable limits.

Third, the trade-off relationship among the simulated fluxes implicit in the results of multivariate calibration studies may be a consequence of deficiencies in model structure and parameterizations (Fenicia et al., 2007; Hogue et al., 2006). However, in the calibration strategies adopted in most studies, including the assumption of a complementary relationship among the fluxes, combining objectives and lack of a definition of error thresholds prevent any meaningful diagnoses of the model. For example, the limitations of the stepwise calibration methodology for identifying deficiencies in model structure and parameterizations is well known (Fenicia et al., 2007). Additionally, most multivariate calibration studies are

deterministic and hence are inappropriate for studying the dierences in optimal parameter sets between univariate and multivariate calibration cases, as they do not address the issue of equifinality (Beven, 1996, 2001). Even in studies that use stochastic methods such as ensemble Kalman filter (Wanders et al., 2014), there is little discussion on how parameters behave between different calibration cases.

In this study, we combine a formal Bayesian calibration approach with the concept of Pareto optimality to address the issues detailed above. We utilize a formal Bayesian calibration approach to define the limits of acceptability or error thresholds in order to distinguish between behavioral and non-behavioral solutions (Beven, 2006; Vrugt et al., 2009b) for each of the incorporated water balance components. Behavioral solutions are model parameter sets that result in errors that are within a defined threshold or limit with respect to a specific simulated response (for example ET or SM). If a trade-off relationship does exist among the incorporated fluxes, as opposed to a complementary relationship, the concept of Pareto optimality would help in understanding the extent to which the accuracy of a particular flux can be improved without affecting, to an unreasonable degree, the accuracy of other fluxes. In addition, Pareto optimal solutions are unbiased by any subjective weights given to any particular flux or storage component over another, unlike simultaneous calibration strategies (Gupta et al., 1998). Hence, we use Pareto optimality-based calibration to create a set of non-dominated solutions that characterize the trade-offs among the incorporated variables. We develop a multivariate calibration framework that combines behavioral solutions from Bayesian calibration and multivariate calibration solutions to address the following research questions: 1) Does incorporation of multiple fluxes into calibration produce parameter distributions that are behavioral with respect to all fluxes considered for calibration? 2) For a given large scale hydrologic model, what is the extent of trade-off, if any, in accurate simulations of multiple fluxes considered for calibration? 3) Can behavioral and multivariate calibration solutions help identify deficiencies in hydrologic model parameterization that lead to trade-offs in the accurate simulations of multiple variables?

3.2 Methodology

3.2.1 Conceptual framework

Consider a hydrologic model,

$$O = \mu(\theta, I) \tag{3.1}$$

where O is a matrix consisting of model output or responses (such as evapotranspiration, soil moisture, streamflow etc.), I is a matrix consisting of model input (meteorological forcings such as precipitation, air temperature, etc.), and μ represents the mathematical structure of the hydrologic model, typically a deterministic or stochastic function such that $\mu : I \mapsto O$; θ is a vector of model parameters (Kavetski et al., 2006). Given a matrix of observations, $\hat{O}$, a measure, L, can be defined as

$$L(E(\theta) = O(\theta) - \hat{O}) = S \tag{3.2}$$

where E is the error residual matrix, L is a measure or metric that preserves the information contained in the residuals (such as mean absolute error or root mean square error), and $S \in (-\infty, \infty)$ is some scalar quantity that represents the value of L.

Assuming that the model structure (μ) and the upper and lower limits of the parameters (θ) are fixed, feasible bounds for equation 3.2, termed as objective space, can be defined (Gupta et al., 1998) (a conceptual representation of a feasible objective space for two objectives (L_1 and L_2) is shown in Figure 3.1.). In traditional model calibration, an optimal parameter set, θ^*, is identified by minimizing equation 3.2 ($L_1^* = L(\theta_1^*)$ and $L_2^* = L(\theta_2^*)$ are the optimal values for objective 1 and 2 in Figure 3.1). Using formal or informal Bayesian approaches, it is also possible to identify sets of parameters, θ^b, that result in behavioral solutions, based on a defined cut-off threshold (Vrugt et al., 2009b) ($L_1^b = L(\theta_1^b)$ are behavioral solutions for objective 1, represented by the objective space to the left of cutoff threshold e_1 and $L_2^b = L(\theta_2^b)$ are the behavioral solutions for objective 2, represented by the objective

space below e_2 in Figure 3.1). To quantify the trade-offs between multiple objectives (L_1 and L_2 in this example), the concept of Pareto-optimality can be used. This results in a set of parameters, θ^p, that give rise to non-dominated or Pareto-optimal solutions for the objectives considered ($L^p_{1,2} = L(\theta^p_{1,2})$ are the Pareto-optimal solutions for two objectives, represented by the points along the red line in Figure 3.1).

In this study, the measure L the is root mean square error (RMSE) and the model responses, O, considered for calibration are evapotranspiration (ET), soil moisture (SM) and streamflow (SF). To address the first research question, we test the following hypothesis:

$$L^{p,b}_{1,2,..n} = L^p_{1,2,..n} \cap (L^b_1 \cap L^b_2 ... \cap L^b_n) \neq \emptyset \tag{3.3}$$

where $L^p_{1,2,..n}$ is the non-dominated or Pareto-optimal solutions for n objectives, ($L^b_1 \cap L^b_2 ... \cap L^b_n$) is the intersection of behavioral solutions for n objectives (blue space in Figure 3.1 for two objectives), $L^{p,b}_{1,2,..n}$ are solutions that are both non-dominated and behavioral with respect to the n-objectives (blue points in Figure 3.1) and $\emptyset$ represents an empty set. In other words, for incorporation of multiple fluxes in hydrologic model calibration to be considered valuable, it should be possible to identify a set a of parameters that result in both Pareto-optimal and behavioral solutions for all the fluxes considered in calibration. We note that in defining the hypothesis, we have considered only non-dominated or Pareto-optimal solutions in the multi-objective space (blue points in Figure 3.1). However, any solution within the behavioral limits, non-dominated or dominated, can be considered as valid multivariate calibration solution. In such cases, the hypothesis to be tested reduces to $L^b_1 \cap L^b_2 ... \cap L^b_n \neq \emptyset$. In other words, the intersection of behavioral solutions for n objectives in the multi-objective space (blue space in Figure 3.1) must be a non-empty set.

In this study, we only focus on the Pareto optimal solutions to test the hypothesis (Equation 3.3) and analyze the trade-off among accurate simulations of multiple fluxes. In addition, the behavioral solutions from Bayesian calibration and the Pareto optimal solutions are discrete (unlike the representation in Figure 3.1). Therefore, the probability of finding an intersection between the two is extremely low. Hence, the non-dominated solutions, $L^p_{1,2,..n}$,

that lie within the error thresholds or limits of acceptability $\{e_1, e_2, ...e_n\}$ are considered as $L^{p,b}_{1,2,..n}$ (Figure 3.1 provides an example for two objectives). We test the above hypothesis for different pairs of model responses ($n = 2$): 1) evapotranspiration and soil moisture (ET-SM), 2) evapotranspiration and streamflow (ET-SF), and 3) soil moisture and streamflow (SM-SF).

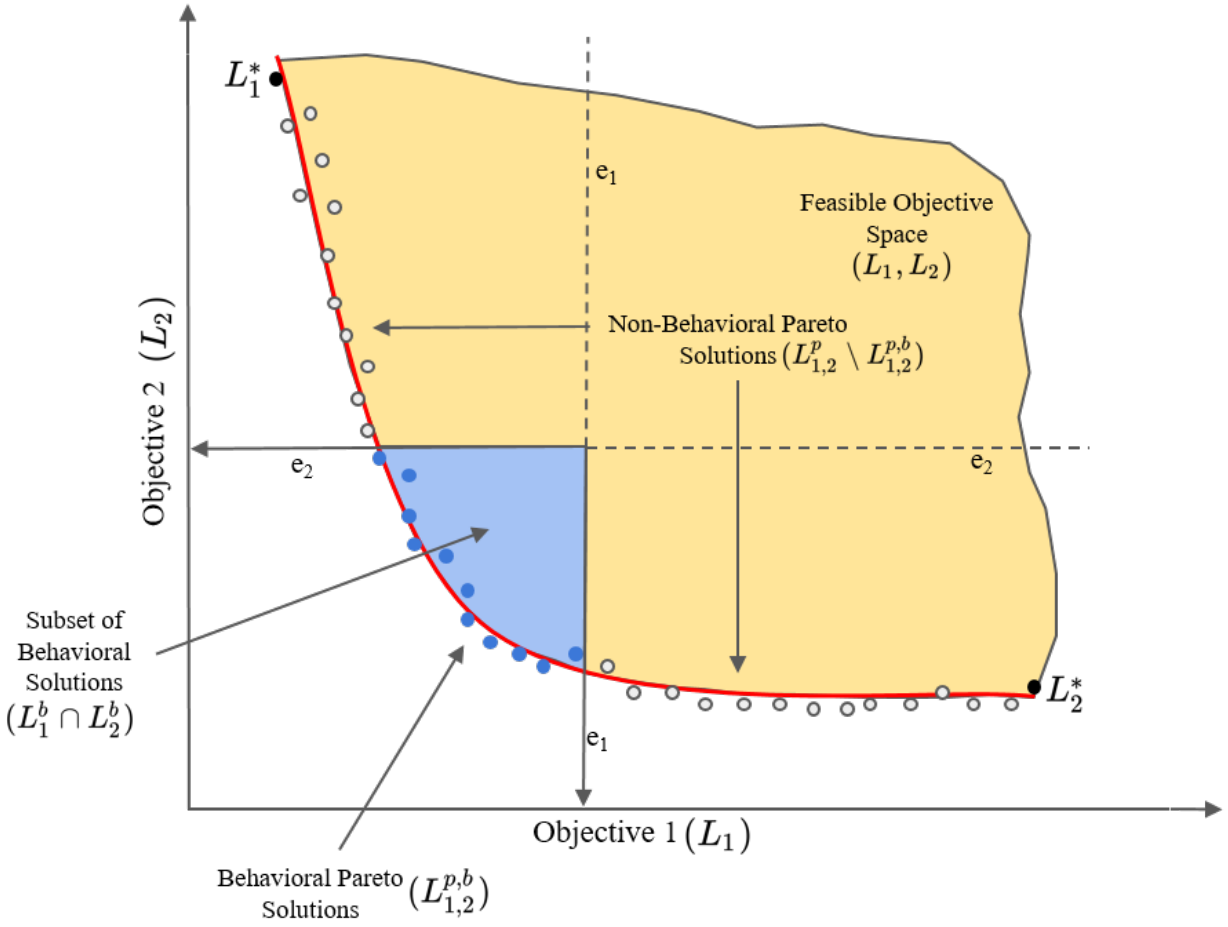

Figure 3.1: Conceptual framework of the methodology adopted in this study (adapted from Efstratiadis and Koutsoyiannis (2010)).

3.2.2 Defining limits of acceptability for individual model responses

To define the limits of acceptability, $e1$ and $e2$, we adopt a formal Bayesian approach to derive the posterior distribution of model error (RMSE) and parameters for individual model

responses (ET, SM, and SF). Specifically, we utilize the Differential Evolution Adaptive Metropolis (DREAM) Markov Chain Monte Carlo (MCMC) scheme (Vrugt et al., 2009a, 2008) that has been applied to Bayesian or uncertainty-based calibration of hydrologic (Shafii et al., 2014) and hydrogeologic (Laloy et al., 2013) models. There is a specific reason for using DREAM in this study. Unlike informal Bayesian approaches such as Generalized Likelihood Uncertainty Estimation (GLUE) (Beven and Binley, 1992), the definition of cutoff thresholds to distinguish behavioral solutions is not subjective. Instead of rejecting solutions based on a subjectively defined threshold, DREAM uses a formal likelihood function to assign probabilities to all the solutions that form the final converged posterior distribution. As a result, a specific quantile of the sampled probability distribution can be considered the cutoff threshold for distinguishing behavioral solutions (Vrugt et al., 2009b).

In this study, we assume no apriori knowledge about the value of error (RMSE) thresholds for any of the model responses (ET, SM, and SF). Instead, we determine a set of limits of acceptability corresponding to 10%, 25%, 50%, 75%, 90%, 95% and 99%, quantiles from the posterior distribution of RMSE, derived using DREAM for ET, SM, and SF. Note that the likelihood function in DREAM considers error residuals and not RMSE to determine the posterior distribution of parameters. Recent advances such as approximate Bayesian computation has enabled the use of summary statistics and error metrics (such as RMSE) for diagnostic model calibration (Vrugt and Sadegh, 2013; Sadegh and Vrugt, 2014) and evaluation (Gupta et al., 2008). We do not use these methods in this study, as they require apriori definition of the limits of acceptability, similar to GLUE. We note that the DREAM solutions in this study are derived using residual and likelihood-based fitting methods, whereas the Pareto optimal solutions are derived using RMSE as the objective function (described below).

3.2.3 Pareto optimal solutions for combination of model responses

The value of using the concept of Pareto optimality for calibration of hydrologic models is well documented (Gupta et al., 1998). Studies have focused on identifying the best objec-

tive functions for improving streamflow calibration (Efstratiadis and Koutsoyiannis, 2010), developing systematic multi-objective calibration frameworks (Madsen, 2003) and analyzing the resulting Pareto fronts (Khu and Madsen, 2005). Multi-objective methods have been applied at small scales to constrain land surface model parameters (Gupta et al., 1999) and evaluate model performance and parameter behavior (Hogue et al., 2006) using multiple fluxes. However, the utility of such an approach for multivariate calibration and diagnosis of large scale hydrologic models has received relatively less attention. In this study, we use A Multi-Algorithm Genetically Adaptive Multiobjective (AMALGAM) algorithm (Vrugt and Robinson, 2007) and RMSE as the objective function to derive non-dominated solutions for the following combinations of model responses: 1) ET and SM (ET-SM), 2) ET and SF (ET-SF), and 3) SM and SF (SM-SF).

3.2.4 Hypothesis testing, trade-off analysis, and model diagnosis

To understand whether the models can accurately simulate multiple water balance components, we combine the behavioral solutions from DREAM and the Pareto optimal solutions from AMALGAM as detailed above. Specifically, we test the hypothesis defined in Equation 3; the set of non-dominated solutions derived for different combinations of fluxes (ET-SM, ET-SF, and SM-SF) is a non-empty set for a particular behavioral limit (10%, 25%, 50%, 75%, 90%, 95% and 99% quantiles from the posterior distribution of RMSE). If a particular combination of fluxes has at least one Pareto optimal point within a stricter definition of error threshold compared to another combination, then the model is better at simulating the former combination of fluxes together. For example, consider that the ET-SM combination has at least one Pareto optimal solution within the 50% quantile error thresholds. On the other hand, consider that the ET-SF combination has at least one Pareto optimal solution within the 25% quantile error thresholds. In such a case, it can be concluded that it is valuable to incorporate ET and SF together compared to ET and SM.

To study the trade-off in the accuracy of the simulated model responses, we analyze the Pareto front qualitatively and quantitatively. Qualitatively, a well-defined Pareto front

implies that the incorporated model fluxes exhibit a trade-off relationship as opposed to a complementary relationship. Quantitatively, the range of the objectives and the slope of the Pareto-front can help compare different Pareto fronts in order to understand the trade-off relationship between the different combinations of fluxes. Specifically, we define magnitude of trade-off as the increase in the error of a specific model response required to affect a unit decrease in the error of the additional flux. We calculate and compare the average, maximum, and minimum magnitude of trade-off for each of the three multivariate calibration cases from the slopes of the Pareto front.

To diagnose the reasons for trade-offs in accuracy among different model responses, we study the differences in model parameter distributions among 1) behavioral solutions of individual responses, 2) behavioral solutions of individual model responses and behavioral Pareto optimal solutions, and 3) behavioral solutions of individual responses and Pareto optimal solutions that are not behavioral for any model response. We compare the empirical cumulative distribution functions (ECDFs) and quantiles of parameters to help identify parameters that most affect model behavior when additional fluxes are incorporated. We quantify the difference between the PDFs of the parameters using Hellingers distance, H, a statistical distance measure defined as

$$H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}(\sqrt{p_i} - \sqrt{q_i})^2} \tag{3.4}$$

where, $P = (p_1, p_2, ...p_k)$ and $Q = (q_1, q_2, ...q_k)$ are discrete probability distributions. We determine the Kolmogorov-Smirnov (KS) test statistic to compare the ECDFs of different parameter distributions. The KS test statistic determines the maximum distance between two ECDFs. In addition, we examine how well the objective functions in the Pareto optimal solutions are correlated with the corresponding parameter sets. This will help us map the changes in parameter values along the Pareto front and identify parameters that influence the trade-off relationship between the objectives.

3.3 Experiment design

3.3.1 Study area and time period

To simulate large-scale hydrologic studies, we choose the Mississippi basin in the United States as the study region. The basin covers an area of about 3.3 million sq. km. and six USGS HUC-2 (Hydrologic Unit Code) basins (Figure 3.2). The average temperature over the basin is about 12^{o}C and the annual average rainfall is estimated as 800mm (Cai et al., 2014). Cai et al. (2014) also classify the Ohio and Tennessee regions as wet regions, the Missouri basin as dry and the Upper Mississippi region as a transitional region between wet and dry. We use historical data from the year 2004 for calibration and data from the year 2005 for validation. Employing a monthly time step, 72 streamflow data points and around 63000 ET and SM data points are used for calibration. The reasons reasons for selecting a single year for calibration and validation of the model are two-fold: 1) The study is a calibration experiment that does not seek to produce the best-performing hydrologic model for the Mississippi Basin. Rather, the primary aim is to rigorously test whether large-scale hydrologic models can behaviorally simulate multiple fluxes and study the reasons behind the trade-offs, if any, between accurate simulation of multiple fluxes. 2) The forward hydrologic model used in the study is computationally expensive.

3.3.2 Observational data

To simulate studies that use a sparse network of streamflow gauges for the calibration of hydrologic models, we use the computed monthly runoff for the six HUC-2 basins sourced from USGS. For calibrating the hydrologic model with remotely sensed ET, we use monthly estimates from the Global Land Evaporation Amsterdam Model (GLEAM) (Martens et al., 2016). We select GLEAM ET based on the findings of Koppa and Gebremichael (2017) in which GLEAM, AVHRR and MODIS ET datasets were ranked using a framework based on the Budyko hypothesis, a semi-empirical model that describes long-term water and energy balance of catchments (Budyko, 1974). The spatial resolution of the GLEAM dataset is

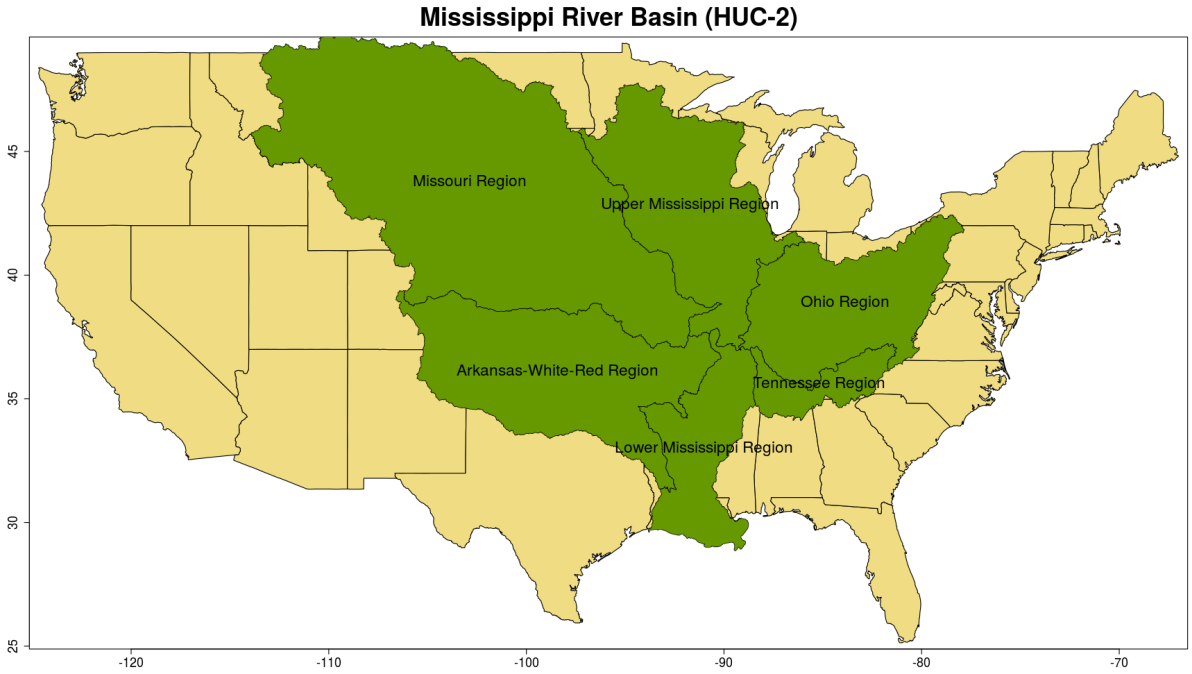

Figure 3.2: A map of the Mississippi basin showing the six USGS HUC-2 basins.

0.25^o x 0.25^o. We use monthly soil moisture estimates from ESA-CCI (Dorigo et al., 2017) for calibrating the hydrologic model with SM. The ESA-CCI dataset is selected based on the availability of data for the study period. The spatial resolution of the ESA-CCI dataset is 0.25^o x 0.25^o. We note that the ESA-CCI soil moisture measurements correspond to the top 5 cm of the soil layer, but the top soil layer of the hydrologic model (detailed below) is 10 cm. Although the unit of measurement is m^3/m^3, this difference in soil layer depths may lead to systematic bias and needs to be corrected. We adjust the values of simulated soil moisture to match the statistics of the observed dataset based on López López et al. (2017) as

$$SM'_{sim} = \frac{\sigma_{SM_{obs}}}{\sigma_{SM_{sim}}} * (SM_{sim} - \overline{SM_{sim}}) + \overline{SM_{obs}} \tag{3.5}$$

where SM'_{sim} is the scaled simulated soil moisture, $\sigma_{SM_{obs}}$ and $\sigma_{SM_{sim}}$ are the standard deviations of the observed and simulated soil moisture, SM_{sim} is the simulated soil moisture to be scaled, $\overline{SM_{obs}}$, and $\overline{SM_{sim}}$ are the means of the observed and simulated soil moisture.

Remotely sensed hydrologic fluxes such as precipitation, ET and SM are subject to large uncertainties due to differences in retrieval algorithms and sensors (Kidd and Huffman, 2011; Gebregiorgis and Hossain, 2014), thus necessitating the need for validating the chosen ET and SM datasets with ground-based measurements. We compare the GLEAM estimates with ground-based flux tower measurements from the Ameriflux network (https://ameriflux.lbl.gov/). A scatter plot of GLEAM versus Ameriflux ET for 2004 shows that GLEAM is capable of representing the ET flux in all the sub-basins to a fair degree of accuracy (Figure 3.3a). The RMSE of GLEAM data is estimated to be 21.4 mm/month. For validating ESA-CCI SM data, we make use of near-surface soil moisture measurements from the TAMU North American Soil Moisture Database (NASMDB) (Quiring et al., 2016). For the study period, SM sensors from only three of the six sub-basins are available. The scatter plot (Figure 3.3b) shows that remote sensing data overestimates the observed soil moisture in the Lower Mississippi Region. The RMSE value for ESA-CCI SM is 0.12 m3/m3.

As these satellite-based datasets are used together in multivariate calibration, we quantify the error in the closure of water balance (Figure 3.3c and 3.3d). First, we compare the annual ET over the six HUC-2 catchments in the study area with the difference between precipitation (P) and runoff (Q) for the years 2000-2009. From Figure 3c, it is evident that the errors in the closure of water balance are quite low for most of the catchments. The exceptions are three years in the Lower Mississippi and Tennessee regions in which inter-annual storage (soil moisture and groundwater) changes may play an important role. Importantly, the water balance closure errors for the calibration and validation periods are low across all the regions, including the Lower Mississippi and Tennessee regions. The mean annual water balance closure error from observational datasets (P - ET - Q), averaged over the entire Mississippi basin, is about 108 mm/year (9 mm/month). When the water balance components are summed over the entire basin, the water balance closure error is approximately 640 mm (53 mm/month). We also make sure that the observational datasets do not exceed catchment-scale water and energy limits as described by the Budyko hypothesis. In this study we make use of Fus equation [Fu, 1981], a single parameter Budyko function that relates the

evaporative index (E/P) and the aridity index (Ep/P, Ep is potential ET) as

$$\frac{E}{P} = 1 + \frac{E_p}{P} - \left(1 + \left(\frac{E_p}{P}\right)^{\omega}\right)^{\frac{1}{\omega}} \tag{3.6}$$

where ω is the Budyko parameter that has no analytic solution. We use a generally accepted representative value of 2.6 in equation 3.6 to construct the Budyko curve (red line in Figure 3.3d). It reveals that catchments are closely clustered around the Budyko curve, except for the Missouri region, and are within the energy and water limits (dotted lines in Figure 3.3d).

3.3.3 Setup and validation of the hydrologic model

To replicate studies that use spatially distributed models, we choose the Noah-MP (Multi-Parameterization) Land Surface Model (LSM) (Niu et al., 2011), driven through NASAs Land Information System (LIS) (Kumar et al., 2006). The Noah-MP model builds on the original Noah LSM by incorporating a dynamic groundwater model, improved representation of vegetation canopy and snow pack. Cai et al. (2014) provide a detailed description and a comprehensive evaluation of the model over the Mississippi river basin. All the static input datasets required for running the Noah-MP model are sourced from NASAs LIS data portal (https://portal.nccs.nasa.gov/lisdata). The important static input datasets are the land cover map, sourced from USGS; the soil texture map from STATSGO, sourced from USDA; and the elevation map from GTOPO30, sourced from USGS. Albedo, greenness fraction and temperature are sourced from NCEP reanalysis. The meteorological forcings required by the Noah-MP model include precipitation, air temperature, surface pressure, specific humidity, wind speed, and radiation. All meteorological forcings are derived from Global Data Assimilation System (GDAS) from the Environmental Modeling Center (EMC) of the National Center for Environment Protection (NCEP) (Derber et al., 1991). The spatial resolution of the dataset is 0.47^o x 0.47^o. The meteorological inputs are interpolated onto the model grid using bilinear interpolation. To minimize the adverse effects of mismatch in the spatial resolution of the model and observations (Samaniego et al., 2010, 2017), the

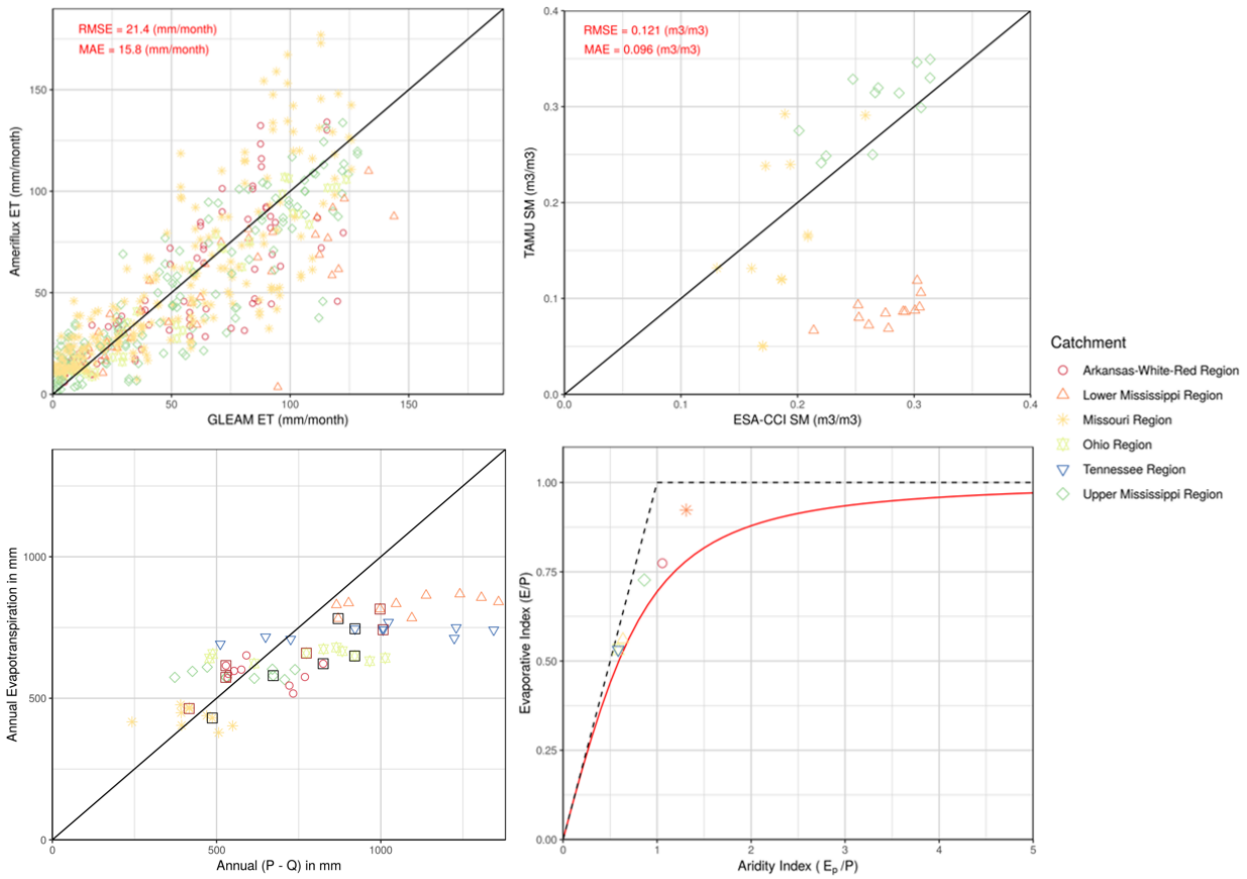

Figure 3.3: Scatter plots of a) GLEAM ET vs Ameriflux measurements (top right panel); b) ESA-CCI soil moisture vs TAMU NASMDB measurements (bottom panel) for 2004; c) annual GLEAM ET vs Annual Precipitation (P) - Runoff (Q) for the years 2000-2009 (bottom left). The errors for the calibration and validation years (2004 and 2005) are highlighted by black and brown bounding boxes respectively; and d) the Budyko space (Evaporative index vs Aridity index) averaged over the years 2000-2009 for the six catchments. The red line is the ideal catchment water-energy balance as represented by the Budyko hypothesis. The dotted lines represent the water (horizontal line) and energy (diagonal line) limits (bottom right).

Noah-MP model is set-up for the Mississippi river basin at a spatial resolution of 0.25^o x 0.25^o (similar to the resolution of GLEAM ET and ESA-CCI SM). The Noah-MP model is spun-up for a period of 68 years by looping through the year 2003 until the groundwater and

soil moisture storage reach equilibrium. The model time step is three hours. The number of soil layers in the model is four with thicknesses 10cm, 30cm, 60cm and 100cm. Specific Noah-MP model physics options selected for dierent processes are detailed in Table 3.1.

The Noah-MP model contains 71 standard parameters (present in user-defined tables) and 139 hard-coded parameters (present in the model code). The Noah-MP model output has been found to be sensitive to about two-thirds of the 71 standard parameters (Cuntz et al., 2016). As the study is a calibration experiment involving multiple calibration cases, we keep the parameter dimension of the calibration problem manageable by selecting five of the most sensitive parameters from the Cuntz et al. (2016) study. The selected parameters are two surface runoff-related parameters (REFDK and REFKDT), the exponent in the Brooks-Corey equation (BB), soil porosity (MAXSMC), and hydraulic conductivity at saturation (SATDK). Of the five parameters, BB, MAXSMC, and SATDK are related to soil texture. As there are twelve soil texture classes, the total number of parameters selected for calibration in the Noah-MP hydrologic model is 38 (Table 3.2 presents a detailed breakdown of the parameters with maximum and minimum values used for calibration). We select the minimum and maximum values of the parameters from literature (MAXSMC and SATDK values from Cai et al. (2014), BB and REFDK values from Cosby et al. (1984), and REFKDT values from Mendoza et al. (2015)). We adjust the minimum and maximum values to improve the rate of convergence of the calibration algorithms.

To ascertain whether the Noah-MP model and the parameters considered for calibration can simulate ET, SM, and SF accurately, we validate the model for the year 2005. For this, we select a parameter set from the DREAM solutions that results in the lowest RMSE value for each of the model responses (ET, SM, and SF). We present a time series comparison of observed and simulated monthly ET, SM, and SF for the six HUC-2 sub catchments (Figure 3.4a). When the Noah-MP model is calibrated with GLEAM ET (top panel), it is evident that the model performs very well in simulating the observed ET values and seasonality. One exception is the underestimation of ET in the summer months for all regions except the Ohio region. The results are consistent across the six HUC-2 hydrologic regions in the

calibration (first 12 data points) and validation time periods (remaining 12 data points). The close match between modeled and observed ET is reflected in the scatter plots of the annual totals of ET as well (Figure 3.4b). Relatively higher variance is observed in the SM results when the Noah-MP model is calibrated with ESA-CCI soil moisture (middle panel). The simulated soil moisture for all the regions is quantitatively consistent with the observed SM values for the top soil layer. However, we see some discrepancy in the seasonality of soil moisture between the simulated and observed soil moisture; it is especially pronounced in the last six months of 2005 in the Tennessee region (middle panel, fifth column) and the first few months of 2004 in the Arkansas-White-Region (middle panel, first column). In the Missouri region (middle panel, third column), the model simulates the observed seasonality but is unable to capture the peaks in the observed SM perfectly. In the Upper Mississippi region (middle panel, sixth column), the model has difficulty in simulating the timing of the troughs (May and June of 2004 and 2005) seen in the observed SM. Similar to the ET results, the annual average SM consistently matches the observed values for all six HUC-2 basins (Figure 3.4b). The results for streamflow simulated by the SF-calibrated Noah-MP model present a more consistent picture (bottom panel). They show that the model generally performs well for all six HUC-2 catchments. However, similar to soil moisture results, there are some inconsistencies in simulating the seasonality of the last six months of 2005, especially in the Arkansas-White-Red (bottom panel, first column), the Ohio (bottom panel, fourth column) and the Upper Mississippi (bottom panel, sixth column) regions. The model also performs well in the Lower Mississippi and Missouri regions, but the peaks are higher in 2004 (June and July) compared with the observed streamflow time series. At annual timescales, the SF-calibrated Noah-MP model overestimates SF for the year 2005 in the Ohio basin (June to October 2005). We see that the errors in evapotranspiration and streamflow are comparable to the results of the Ma et al. (2017) study. For the six HUC-2 sub-basins considered in this study, Ma et al. (2017) report a RMSE of about 10 mm/month, whereas the RMSE in this study is about 18mm/month. The average RMSE of ET, calculated over the entire US by Ma et al. [2017], is about 10 mm/month, which matches the RMSE of the time series of ET presented in Figure 4 (about 10mm/month). We note that the Ma et al. (2017) study does

not calibrate the Noah-MP model.

Table 3.1: Noah-MP model physics options

Model Physics	**Selected Physics Option**
Vegetation model	Use table Leaf Area Index (4)
Canopy stomatal resistance	Ball-Berry (1) (Ball et al., 1987)
Soil moisture factor for stomatal resistance	Original Noah (1) (Chen et al., 1997)
Runoff and groundwater	TOPMODEL with groundwater (1) (Niu et al., 2007)
Surface layer drag coefficient	Original Noah (2) (Chen et al., 1997)
Frozen soil permeability	Linear effects, more permeable (1) (Niu and Yang, 2006)
Radiation transfer	Modified two-stream (1) (Yang and Friedl, 2003)
Snow surface albedo	CLASS (2) (Verseghy et al., 1991)
Rainfall and snowfall Partitioning	Jordan Scheme(1) (Jordan, 1991)
Lower boundary of soil temperature	Original Noah (2) (Chen et al., 1997)
Snow and soil temperature time scheme	Semi-implicit (1)
Super-cooled liquid water	No iteration (1) (Niu and Yang, 2006)

[a] The number in the brackets represents the internal Noah-MP model code for the selected physics option

3.3.4 Setup of DREAM and AMALGAM algorithms

DREAM is a multi-chain Markov chain Monte Carlo (MCMC) simulation algorithm that automatically tunes the scale and orientation of the proposal distribution en route to the target distribution. It is designed for increasing the sampling efficiency of complex, high-dimensional parameter spaces, while maintaining detailed balance and ergodicity (Vrugt, 2016). In this study, we use the MT-DREAM (ZS) version of DREAM, which utilizes multi-

Table 3.2: Details of Noah-MP parameters for calibration

Parameter	**Total Parameters**	**Units**	**Minimum**	**Maximum**
REFDK	1	m/s	1.4e-06	6.5e-06
REFKDT	1	No Units	1.0	5.0
BB1 - BB12	12	No Units	0.5	12.0
MAXSMC1 - MAXSMC12	12	No units	0.1	0.7
SATDK1 - SATDK12	12	m/s	2.0e-06	7.03-02

[a] Soil texture classes for BB, MAXSMC and SATDK (from 1 - 12):
Sand, Loamy sand, Sandy loam, Silt loam, Silt, Loam, Sandy clay loam, Silt clay loam, Clay loam, Sandy clay, Silty clay and Clay.

try sampling (MT), snooker updating and sampling from an archive of past states to improve the rate of convergence and make use of parallel computing resources. Specific configuration options and parameters of the MT-DREAM (ZS) algorithm used in this study are detailed in Table 3.3. We select the Laplacian likelihood based on the findings of Schoups and Vrugt (2010); residual errors in rainfall-runoff models of humid basins, like the Mississippi basin (In Figure 3.4d, most basins are within aridity index of 1.0), are better represented by a Laplacian distribution than a Gaussian distribution. The likelihood function is used to summarize the distance between the model simulations and the corresponding observations. For ET and SM variables, error residuals determined at all 0.25^o x 0.25^o grid cells and time steps (monthly) across the entire Mississippi river basin (all six HUC-2 basins together) are used to determine the likelihood function. Similarly, SF error residuals are determined using simulated and observed runoff at all six HUC-2 basins and all months of the calibration period. On a workstation with 16 processors, MT-DREAM (ZS) required approximately 16 days (14500 iterations) to converge to a solution for each of the model responses (ET, SM, and SF), with each iteration of the Noah-MP model taking around 20 minutes to complete.

The multi-objective calibration algorithm, AMALGAM, combines the strengths of multiple evolutionary algorithms to improve the speed and efficiency of finding the Pareto op-

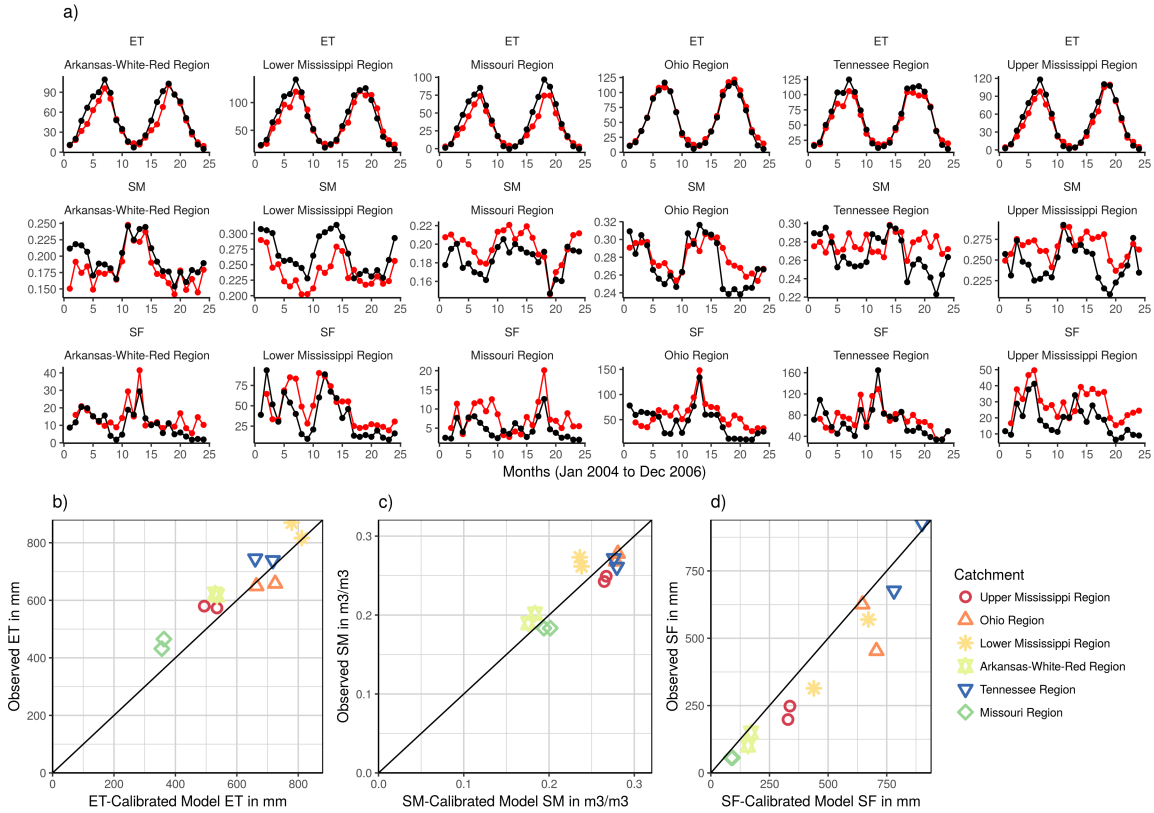

Figure 3.4: a) Time series plots of a) ET-calibrated model ET (red) and GLEAM observed ET (black) in mm/month (Top panel), b) SM-calibrated model SM (red) and ESA-CCI observed SM (black) in m^3/m^3 (Middle panel) and c) SF-calibrated model SF (red) and observed HUC-2 runoff (black) in mm/month (Bottom panel) for the six HUC-2 sub-catchments of the Mississippi basin. The first 12 months correspond to the calibration period of 2004 and the next 24 months correspond to validation years 2005; and b) Scatter plot of annual modeled ET, SM and SF and observed values for the calibration and validation years over the six HUC-2 basins.

timal solutions for multi-objective optimization problems (Vrugt and Robinson, 2007). In the current implementation, four search algorithms are run simultaneously in AMALGAM: differential evolution (Storn and Price, 1997), particle swarm optimization (Kennedy and Eberhart, 2001), adaptive Metropolis (Haario et al., 2001), and NSGA-II (Deb et al., 2002). In AMALGAM, offspring creation is adaptive; the best performing algorithms in the present generation are weighted more in the creation of offspring for the next generation. Specific

configuration options and parameter values of the AMALGAM algorithm used in this study are detailed in Table 3.3. As stated in the methodology section, the objective function is the root mean square error metric. For ET and SM, RMSE is estimated by calculating the error between the modeled and the observed quantities at each 0.25^o x 0.25^o grid cell inside the entire Mississippi river basin and at each time step (each month of the year 2004). For SF, error residuals calculated for all six HUC-2 basins and twelve months are used for estimating RMSE. In other words, the calibration of the Noah-MP model is carried out for the entire Mississippi river basin using a single objective function (Laplacian likelihood in the case of DREAM and RMSE in the case of AMALGAM) and not for the individual HUC-2 basins. On a 16-processor workstation, each of the three multi-objective calibration scenarios (ET and SM, ET and SF, SM and SF) required around 10 days (9000 iterations) to arrive at the final Pareto front.

Table 3.3: MT-DREAM (ZS) and AMALGAM configuration

DREAM Option	**Specified Option**
Number of generations	600
Number of Markov chains	3
Number of forward model parameters	38
Number of crossover values	3
Number of Multi-tries	4
Number of chain pairs proposal	1
Likelihood function	Laplacian likelihood
AMALGAM Option	
Population size	150
Number of generations	60
Number of objective functions	2
Sampling strategy	Latin Hypercube

[a] Note: All other MT-DREAM (ZS) and AMALGAM parameters are set to default values

3.4 Results and discussions

To help understand the impact of calibrating the Noah-MP model with a single water balance component (ET, SM, and SF) on other model responses and to define the limits of acceptability, we first present the results of the posterior distributions of model errors (RMSE) from DREAM. We then test the central hypothesis of this study (equation 3.3) by combining the limits of acceptability and the multivariate calibration solutions from AMALGAM. Next, we address the second research question by analyzing the trade-offs among accurate simulations of ET, SM, and SF as represented by the Pareto fronts. Finally, we show how the developed multivariate calibration framework can help diagnose deficiencies in model parameterization.

3.4.1 Posterior distributions of model errors

We present the posterior distribution of model response errors (RMSE) using probability density and empirical cumulative distribution functions (PDF and ECDF) (Figure 3.5). The PDFs and CDFs of three model responses (ET (top panel), SM (middle panel) and SF (bottom panel)) are constructed for three calibration objectives: ET (green), SM (orange) and SF (blue). First, we analyze the impact of calibrating the Noah-MP hydrologic model with only streamflow (blue) on ET (top panel) and SM (middle panel). Our results confirm the findings of previous studies: calibrating a hydrologic model with only SF adversely affects the accuracy of other model responses (Rakovec et al., 2016a,b). While the 50% quantile of the ET error increases by about 15 mm/month, the SM error increases by around 0.1 m^3/m^3. The same conclusion can be extended for univariate calibration with other variables (ET and SM). For example, when the Noah-MP model is calibrated with SM (orange), the errors in ET (top panel) and SF (bottom panel) are very high. Some results stand-out from Figure 3.5: 1) The error distribution of SF produced with the ET-calibrated model is close to the error distribution of SF from the SF-calibrated model (green and blue plots in bottom panel). This finding reinforces the results of previous studies that use ET for calibration (Immerzeel and Droogers, 2008; López López et al., 2017; Zink et al., 2018). However, the absolute

difference between the 50% quantiles is still around 20 mm/month, indicating a significant impact on SF accuracy due to univariate calibration of the model with ET; 2) Similarly, SF-calibration seems to produce lower errors for SM compared to ET (middle panel) but the absolute errors are still very high; 3) Neither SM nor SF calibration can simulate ET with reasonable accuracy, as seen by the disparate error distributions (top panel). In fact, both SM and SF produce similar error distributions for ET, evident by the ECDFs (blue and orange). For comparison, the RMSE of ET and SF are higher than the mean monthly water balance closure error seen in the observational dataset (about 9 mm/month). Therefore, the RMSE values of ET and SF are not significantly biased by the errors in the observational datasets, as represented by the water balance closure error. We derive the limits of acceptability for ET, SM, and SF from the posterior distributions of RMSE. The error thresholds are defined at 10%, 25%, 50%, 75%, 90%, 95% and 99% quantiles (Table 3.4). The table shows that the range of the quantiles is quite low for ET (about 4 mm/month between the 10% and 95% quantiles) and SM (about 0.02 m3/m3 between the 10% and 95% quantiles). This is reflected in the well-defined posterior distributions of ET and SM errors when the model is calibrated with ET (Figure 3.5, green, top panel) and SM (Figure 3.5, orange, middle panel). In the case of SF, the difference between the 10% and 95% quantiles is about 17 mm/month. This discrepancy maybe due to the higher number of data points available for calibrating ET and SM (spatially distributed observations) compared to SF (point data).

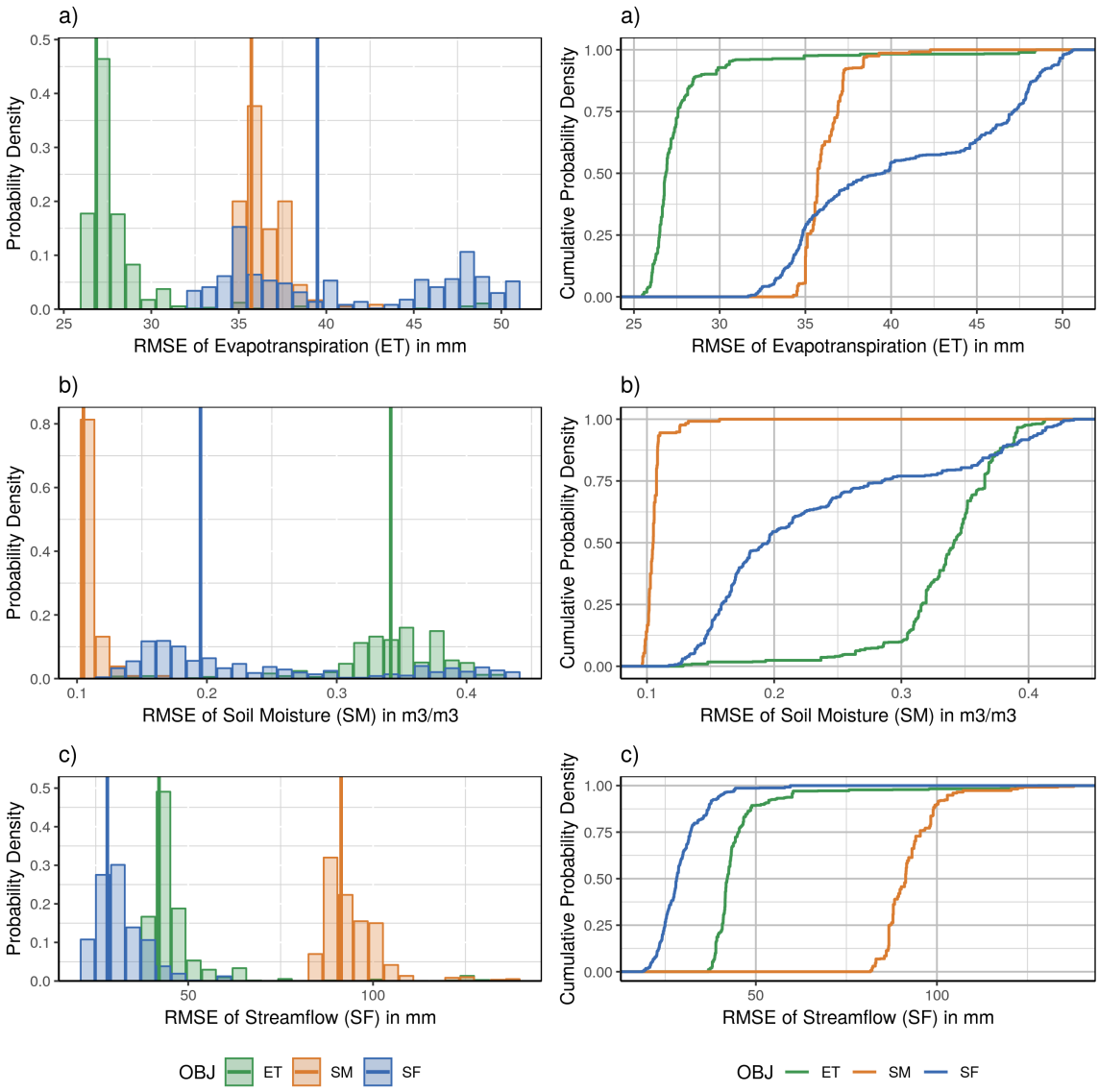

Figure 3.5: A comparison of the posterior probability density functions (PDF) and empirical cumulative distribution functions (ECDF) of root mean square errors of a) evapotranspiration (top panel), b) soil moisture (middle panel) and c) streamflow (bottom panel) when the model is calibrated using DREAM with ET (green), SM (orange) and SF (blue). Vertical lines in the PDFs represent 50% quantiles of RMSE. For comparison, the mean annual water balance closure error from the observational datasets (P - Q - ET) is around 108 mm (9 mm/month). Note: The RMSE values of ET and SM are determined from error residuals calculated at all grid cells of the model domain. The RMSE values of SF are determined using error residuals from all the six HUC-2 regions.

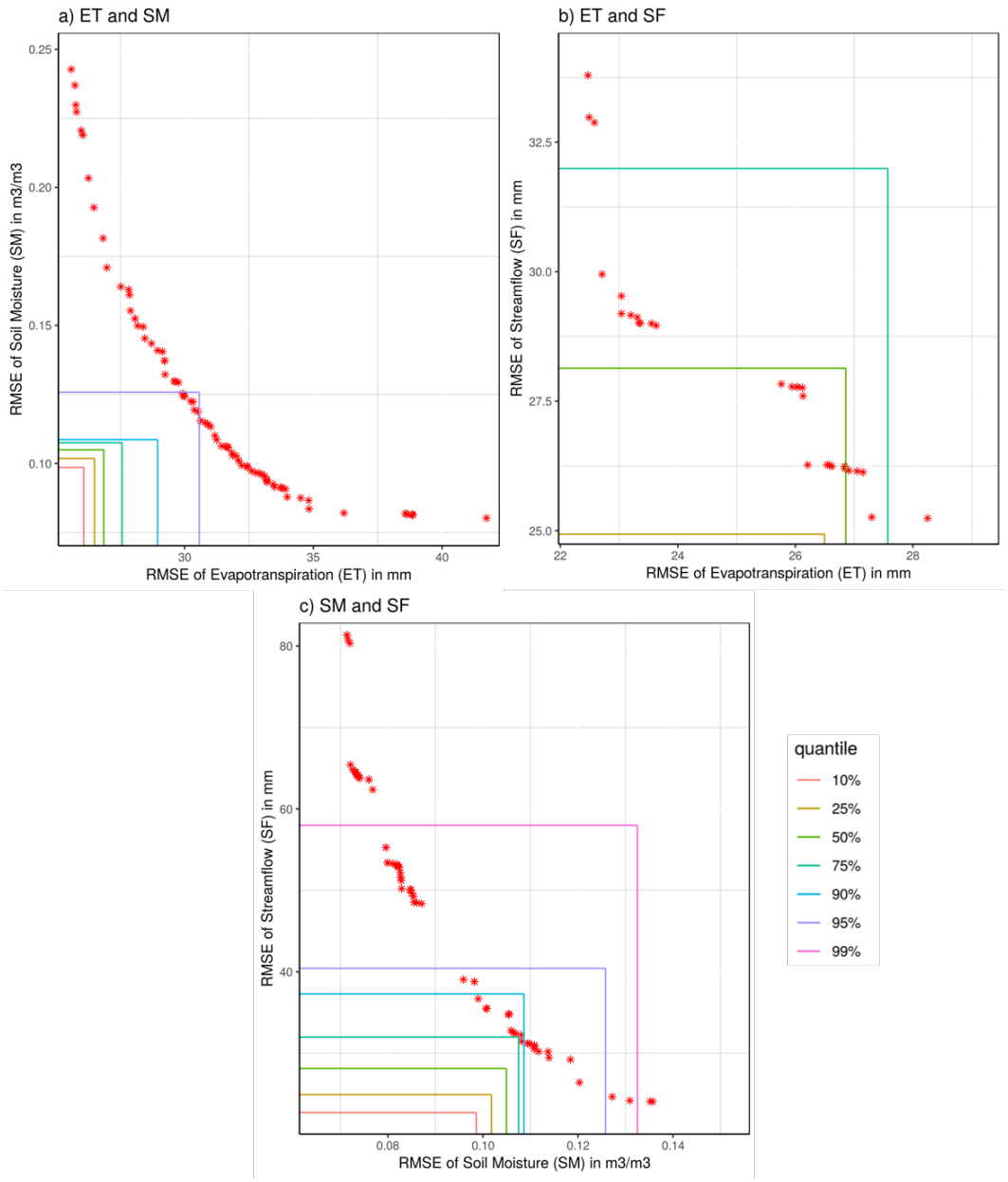

Figure 3.6: Pareto fronts of root mean square errors of a) ET and SM, b) ET and SF and c) SM and SF with limits of acceptability represented by 10%, 25%, 50%, 75%, 90%, 95% and 99% quantiles of the posterior distribution of the error (from DREAM). For comparison, the mean annual water balance closure error from the observational datasets (P - Q - ET) is around 108 mm (9 mm/month). The non-dominated or Pareto optimal solutions are represented by red stars. Note: The RMSE values of ET and SM are determined from error residuals calculated at all grid cells of the model domain. The RMSE values of SF are determined using error residuals from all the six HUC-2 regions.

Table 3.4: Limits of Acceptability for ET, SM and SF derived from posterior distributions of RMSE

Quantile	**ET (in mm)**	**SM (in m^3/m^3)**	**SF (in mm)**
10%	26.1	0.098	22.7
25%	26.5	0.101	24.9
50%	26.9	0.104	28.1
75%	27.6	0.107	32.0
90%	29.0	0.108	37.30
95%	30.6	0.126	40.42
99%	47.9	0.132	58.00

3.4.2 Hypothesis testing using DREAM and AMALGAM solutions

To address the first research question, we test the hypothesis that the intersection of behavioral (from DREAM) and non-dominated solutions (from AMALGAM) is a non-empty set (Equation 3). We present the Pareto fronts of the RMSE of different combinations of model responses (ET-SM, ET-SF, SM-SF) along with different limits of acceptability (Table 43.4) in Figure 3.6. As stated in the methodology section, multivariate calibration with evolutionary algorithms such as AMALGAM may result in Pareto fronts with lesser numbers of non-dominated solutions compared with the initial population size. In this study, we focus only on the Pareto optimal solutions (red stars in Figure 3.6). The non-dominated or Pareto optimal solutions are also used to analyze the Pareto fronts and quantify the trade-offs (next section). We present the number of Pareto optimal solutions of different combinations of model responses that lie within specific RMSE thresholds in Table 3.5.

For the combination of ET and SM model responses (Figure 3.6a), we see that the hypothesis (equation 3.3) fails for all defined error thresholds except for the 95% and 99% quantiles (Table 3.4 for specific values). Even at the 95% quantile, where the ET and SM error thresholds are quite high (30.6 mm/month for ET and 0.126 m3/m3 for SM), only 12

points in the multivariate space can be classified as behavioral. We note that the points near the tails of the Pareto front are well within individual error thresholds of ET and SM. Therefore, we conclude that the Noah-MP model is unable to simultaneously simulate both ET and SM with reasonable accuracy. Even though the incorporation of ET and SM can improve the errors in both the variables, it does not result in improving the realism of the model itself, lending credence to assertion that the relationship between ET and SM may not be complementary. The results are consistent with those of univariate calibration using DREAM, wherein the distributions of the ET- and SM- calibrated model responses have considerable discrepancies between them (compare ET-calibrated ET and SM-calibrated ET (top panel of Figure 3.5) and SM-calibrated SM and ET-calibrated SM (middle panel of Figure 3.5)).

In contrast to ET and SM, the model performs very well in simulating both ET and SF accurately (Figure 3.6b). Even at a stricter error threshold of 50% quantile of individual ET and SF errors (Table 3.4), around 12 points out of the 29 Pareto optimal points can be classified as behavioral (Table 3.5). At the error threshold of the 95% quantile, all 29 Pareto optimal solutions are behavioral. This, along with the fact that the range of errors in the Pareto front is quite small, shows that the relationship between ET and SF can be considered complementary. In other words, the incorporation of ET into SF calibration can improve ET simulation while maintaining the accuracy of SF within reasonable error thresholds. As seen in Figure 3.5, the presence of a complementary relationship between ET and SF is hinted in the posterior distributions (PDF and ECDF) of ET and SF errors (SF-calibrated SF (blue) and ET-calibrated SF (green) in the bottom panel). This result also provides support for studies that incorporate ET and SF into calibration, such as Zink et al. (2018), where incorporation of ET improved ET error by 8% while the NSE of SF reduced by 6% (which could be within the limits of acceptability).

For the combination of SM and SF fluxes, the hypothesis fails for the lower error thresholds (10% to 50% quantiles). At higher quantiles (75% and greater), where the absolute value of the SF error threshold is greater than 32 mm/month (Table 3.4), relatively more

Pareto optimal solutions are classified as behavioral compared to the ET-SM combination. This relative improvement in performance also is reflected in the posterior distribution of individual model response errors (middle panel in Figure 3.5). We see that the error distribution of SF-calibrated SM (blue) is closer to the SM-calibrated SM (orange), compared with the ET-calibrated SM (green). This, combined with the results of the ET-SM Pareto front, shows that incorporation of a storage component such as SM may not lead to improved model performance for multiple fluxes compared with the incorporation of ET. The performance of the SF-SM combination is in line with the findings of **??**, where there are improvements in SF and SM performance when SM is incorporated into calibration. This also highlights the advantages of defining limits of acceptability and casting multivariate calibration as a trade-off problem. Similarly, the results show the drawbacks of assuming a complementary relationship between the model responses. For example, in Rakovec et al. (2016a), the incorporation of TWS improves the RMSE (normalized) of TWS from 0.9 to 0.8 with low impact to SF performance. However, without the specification of an error threshold, it cannot be determined whether the RMSE of 0.8 is behavioral. In other words, the solution may lie on the Pareto front but may still be outside the limits of acceptability. Therefore, the realism of the model may not have improved to a sufficient degree due to the incorporation of an additional flux.

3.4.3 Understanding the trade-offs using Pareto fronts

The results of the hypothesis tests could be a consequence of the nature of trade-offs between the objectives of multivariate calibration, as represented by the Pareto fronts (Figure 3.6). In this section, we address research question 2 and diagnose the results of hypothesis testing. We analyze the characteristics of the trade-offs in accurately simulating the three combinations of model responses incorporated into calibration - 1) ET-SM, 2) ET-SF, and 3) SM-SF. A visual analysis of the Pareto fronts (Figure 3.6) reveals that the ET-SM and SM-SF fronts are relatively well defined compared to the ET-SF Pareto front. The ill-defined ET-SF Pareto front could be an indication that the relationship between ET and SF is complementary, as

Table 3.5: Breakdown of Pareto optimal solutions into behavioral solutions based on different limits of acceptability for ET-SM, ET-SF and SM-SF combinations

Quantile	**ET and SM (N = 123)**	**ET and SF (N = 29)**	**SM and SF (N = 74)**
10%	0	0	0
25%	0	0	0
50%	0	12	0
75%	0	25	0
90%	0	29	12
95%	12	29	27
99%	86	29	58

N represents the total number of Pareto optimal solutions derived from AMALGAM using an initial population of 150

opposed to a trade-off relationship. We compare the number of non-dominated solutions in the multivariate space to test whether these numbers reflect the conclusions of the visual analysis. We see that the ET-SM front has the highest number of non-dominated solutions (123 solutions), followed by SM-SF (74 solutions), and then ET-SF (29 solutions). The higher number of non-dominated or Pareto optimal solutions in the ET-SM front can also indicate a strong trade-off relationship between the ET and SM model responses compared with other combinations. This qualitatively seems to confirm the conclusions drawn from testing the central hypothesis of this study. First, it is consistently more difficult to accurately simulate ET and SM together compared with other combinations. On the other end of the spectrum, the ET and SF fluxes are more complementary to each other, considering 1) the ill-defined shape of the Pareto front and 2) the lesser number (29) of Pareto optimal solutions. In the case of SM and SF model responses (Figure 3.6c), the higher accuracy of the SF-calibrated hydrologic model in simulating SM compared with an ET-calibrated model (middle panel in Figure 3.5) translates to a lesser number of non-dominated solutions compared with the ET

and SM combination.

However, the number of Pareto optimal solutions is not a quantitative measure of the trade-off between the fluxes, as they depend on the optimization algorithm used. Next, we analyze the accuracy trade-offs in simulating the combination of model responses quantitatively. First, we compare the range of the objective functions in each of the three multivariate calibration cases. In the case of ET and SF, the range of the two objectives is very small compared with other combinations. For example, the RMSE of ET flux ranges from 25.5 mm/month to 41.7 mm/month in the ET-SM combination, whereas in the ET-SF combination the range is between 22.5 mm/month and 28.9 mm/month. Similarly, the range of SF in the ET-SF combination (25.2 to 34.1 mm/month) is lower than in the SM-SF combination (24.1 to 81.3 mm/month). Also, the SM range is lower in the SM-SF combination (0.07 to 0.14 m3/m3) compared to the ET-SM combination (0.08 to 0.24 m3/m3). Lower ranges of the objectives in the ET-SF combination imply that both ET and SF can be simulated together to a reasonable degree of accuracy. Lower ranges also may point toward a lower trade-off between two objectives. However, due to differences in the units of the objective functions and dierent ranges across the three Pareto fronts, more analysis is required to draw conclusions.

Finally, we compare the magnitude of trade-offs (defined in the methodology section) across the Pareto fronts of the three calibration cases (ET-SM, ET-SF, and SM-SF). To enable comparison of the trade-offs across the three Pareto-fronts, we normalize the errors in each of the model responses (ET, SM, and SF) by the maximum error, determined over all three combinations. To derive the trade-off matrix (Table 3.6), we fit a second-degree polynomial to the non-dominated points (red stars in Figure 3.6) of the three combinations of model responses - ET-SM (R2 = 0.92), ET-SF (R2 = 0.82), and SM-SF (R2 = 0.96). Next, we divide the Pareto fronts into 15 equally spaced segments. We then calculate the average, maximum and minimum trade-offs from either the slopes or the inverse of the slopes of the 15 segments for each combination of the three model responses. For the combination of ET and SM (first row, second column in Table 3.6), the average increase in the error of ET

required to affect a unit decrease in the error of SM is 2.0 units. For the same improvement, the SF error trade-off is only 0.7 units, implying a larger trade-off in the ET-SM combination compared with the SM-SF combination for soil moisture. Confirming this conclusion, we see that the average trade-o in ET accuracy required to improve SF is much lower than the trade-off required to improve SM. It is interesting to note that there is a higher trade-off in SF accuracy to achieve a unit improvement in ET compared to the trade-o in SM accuracy to achieve the same improvement in ET (first column in Table 3.6). However, as Figure 3.6 shows, more ET-SF multivariate calibration solutions are behavioral compared to ET-SM solutions, and the range of errors is much lower in the ET-SF combination.

Table 3.6: Average, minimum and maximum trade-offs for combinations of ET, SM and SF model responses (minimum and maximum values are within parentheses)

	ET	**SM**	**SF**
ET	-	2.0 (0.5, 7.5)	0.3 (0.08, 0.9)
SM	0.5 (0.1, 2.0)	-	1.0 (0.24, 2.9)
SF	2.0 (1.1, 12)	0.7 (0.35, 4.2)	-

3.4.4 Model diagnosis

Unlike deterministic calibration where a single optimal parameter set is derived, the multivariate calibration framework developed in this study enables the examination of parameter behavior among multiple calibration objectives. Hogue et al. (2006) demonstrated the advantages of Pareto calibration for studying model performance and parameter behavior. In this study, we explore the advantages of combining behavioral solutions from Bayesian calibration and Pareto optimal solutions from multivariate calibration for model diagnosis. The objective is to identify parameters that can explain the significant trade-offs in accuracy among the multiple fluxes incorporated into calibration. Specifically, we try to investigate the reasons behind the higher magnitude of trade-offs in the ET-SM and ET-SF cases. We first compare the posterior probability distributions of calibrated parameters for the univariate

calibration cases (ET, SM, and SF) to identify the parameters that govern the performance of the Noah-MP model with respect to the ET, SM, and SF responses. Figure 3.7 presents a comparison of the ECDFs of the five parameters considered for calibration in this study - REFDK, REFKDT, BB, MAXSMC and SATDK. We note that the ECDFs for parameters BB, MAXSMC and SATDK have been constructed by aggregating parameter solutions from all 12 soil classes (Table 3.2). To keep the analysis consistent with multivariate calibration results (presented later), we only consider behavioral solutions that are within a threshold of the 50% quantile. The ECDFs of the runoff parameters, REFDK and REFKDT, explain the performance of the Noah-MP model with respect to streamflow for different univariate calibration cases. We see that the ECDFs of both REFDK and REFKDT for the SM-calibrated model (orange) are quite different from the SF-calibrated model (blue), leading to relatively poor performance (Figure 3.5). We see that the REFDK parameter distribution from the ET-calibrated model (green) is closer to the REFDK parameter distribution from the SF-calibrated model (blue) compared to the SM-calibrated model (orange). On the other hand, the REFKDT parameter distributions from the ET-calibrated and SM-calibrated models are considerably different from the SF-calibrated model. This shows that the relatively better performance of the ET-calibrated model for the SF flux compared to the SM-calibrated model is more influenced by the REFDK parameter.

We quantify these differences between the parameter PDFs using the Hellingers distance for PDFs (equation 3.4), and the distance between their ECDFs using the Kolmogorov-Smirnov (KS) test statistic. The distance measures are presented in the form of a heat map in which each grid cell represents the statistical distance between the corresponding parameter distributions (Figure 3.8). For example, the first grid cell in the first panel of Figure 3.8a represents the Hellingers distance between the PDFs of the REFDK parameter generated from the ET-calibrated model (first column) and the SF-calibrated model (first row). The three rows in each panel correspond to the univariate calibration objectives (ET, SM, and SF), and the six columns correspond to both the univariate and multivariate calibration objectives (ET-SM, ET-SF, and SM-SF). We see that the Hellingers distance between the

REFDK distributions generated from the ET-calibrated model and the SF-calibrated model (third column and third row) is less than the distance between the distributions generated from the SM-calibrated model and the SF-calibrated model (third column and second row). This difference is more pronounced in the KS statistic heat map (Figure 3.8b). As far as ET is concerned, the inability of both the SM- and SF-calibrated models to accurately simulate ET can be attributed to differences in the posterior distributions of the BB and MAXSMC parameters. As pointed out by Cuntz et al. (2016), the runoff parameters also can influence the ET (and SM) results, as they affect the models water balance. This fact is evident in the case of SM performance, for which the parameter distributions from both the SM-calibrated and SF-calibrated models for BB, MAXSMC and SATDK are similar. However, there is a large difference in the distributions of the runoff parameters, REFDK and REFKDT (Figure 3.8).

The analysis of posterior distributions of parameters from univariate calibration presented above agrees with the results of sensitivity analysis of the Noah-MP model parameters Cuntz et al. (2016). As land surface models are highly complex in terms of parameterization, the same conclusions may not hold when multiple fluxes are incorporated into calibration. We calculate the correlation between the objective functions (RMSE) for the combinations of model responses (ET-SM, ET-SF, and SM-SF) and the corresponding parameters (REFDK, REKDT, BB, MAXSMC, and SATDK). In other words, we map the behavior of the parameters along the Pareto front. From Table 3.7, it is clear that most parameters show opposite correlation for the pair of objectives. For the combination of ET and SM objectives, parameters REFDK and BB show strong correlation with both ET (positive) and SM (negative) errors. This may indicate that these parameters are more responsible for the higher trade-off in accuracy (slope of the Pareto curve) between ET and SM seen in Figure 3.6. Most parameters have almost equal magnitude (but different signs) of correlation with the objective functions, except for REFKDT. In the ET-SF combination, it is clear that the magnitude of correlation between the objective functions and the parameters are significantly lower than the correlations for the ET-SM combination. This clearly reflects the

lesser trade-off and better simultaneous simulations of ET and SF compared with ET and SM. It is interesting to note the higher correlation for the REFDK parameter compared with REFKDT, considering that the difference between the ET-calibrated and SF-calibrated REFKDT parameters was higher than the REFDK parameter. This shows that the parameters that govern model performance for individual fluxes may not correspond to the parameters that affect the trade-off among the combination of fluxes. Also, it highlights the advantage of employing Pareto-based calibration. A similar conclusion can be drawn when SM and SF are combined; we see that the trade-offs are governed by BB and MAXSMC parameters and not the REFDK or REFKDT variables that govern the performance of SF.

Next, we compare the parameter distributions of behavioral multivariate calibration solutions of the ET-SF combination with parameter distributions of models calibrated with only ET and SF. Behavioral solutions with respect to both incorporated model responses are derived at the 50% quantile threshold (Table 3.4). As the number of ET-SF Pareto optimal solutions within the 50% behavioral threshold are few (12), we use both the Pareto optimal and dominated solutions (40) to derive the ECDFs of the five parameters. We note that the forty Pareto optimal and dominated solutions used for deriving the parameter distributions are behavioral (within 50% error quantile), and therefore can be considered as valid multivariate calibration solutions. In addition, we stress here that, unlike DREAM, the parameter distributions derived from multivariate calibration are not true posterior distributions. Therefore, we intend to use the analysis of parameter distributions from only as a tool to investigate difference in parameter values between univariate and multivariate calibration solutions. First, we compare the distance between behavioral ET-SF and SF-calibrated solutions. We see that the parameter distribution of the REFKDT parameter in the ET-SF calibrated model is much closer to the REFKDT distribution in the SF-calibrated model (fifth column and first row in Figure 3.8) compared with the REFKDT distribution in the ET-calibrated model (fifth column and third row in Figure 3.8). However, the Hellingers distance and KS statistic between the ET-calibrated and SF-calibrated models (first column and first row in Figure 3.8) for the REFDK parameter is smaller compared to the distance

between behavioral ET-SF solutions and the SF-calibrated model (fifth column and first row in Figure 3.8). In fact, the incorporation of ET and SF into the calibration only improves the REFKDT parameter distribution, and all other parameters show greater distances compared with parameters derived from models calibrated with only ET or SF. This shows that the surface runoff parameter, REFKDT, is very dominant in governing the behavioral simulation of both the ET and SF fluxes. Finally, we try to determine the reasons for the poor combined simulation of the ET-SM and SM-SF variables. To do this we compare the multivariate calibration solutions from the ET-SM and SM-SF calibration cases that are not behavioral with respect to any of the individual fluxes in models calibrated with only the individual fluxes (ET, SM, and SF). Considering the case of the ET-SM combination, we see that the incorporation of both the ET and SM responses reduce the distances between the distributions of parameters that govern ET (fourth column and third row for BB, MAXSMC and SATDK in Figure 3.8) compared with the SM-calibrated model (fourth column and second row for BB, MAXSMC and SATDK in Figure 3.8). However, the distributions of the REFDK variable, which seem to influence the trade-off in accuracy between the ET and SM variable (Table 3.7), deviate significantly from the SF-calibrated model (compare Hellingers distance between the ET-SM combination and SF, and the ET-SM combination and ET for the REFDK parameter in Figure 3.8). This shows that a surface runoff parameter such as REFDK can influence the combined simulation of ET and SM. We can draw a similar conclusion in the case of SM-SF, where the incorporation of SM and SF improves the performance of the runoff parameters (REFDK and REFKDT) compared with the SM-calibrated model. However, the parameter that seems to influence the trade-off more, MAXSMC, shows degradation compared with the SM-calibrated model. It is interesting to note that the combination of SM and SF adversely affects a soil moisture parameter, even though SM has been incorporated.

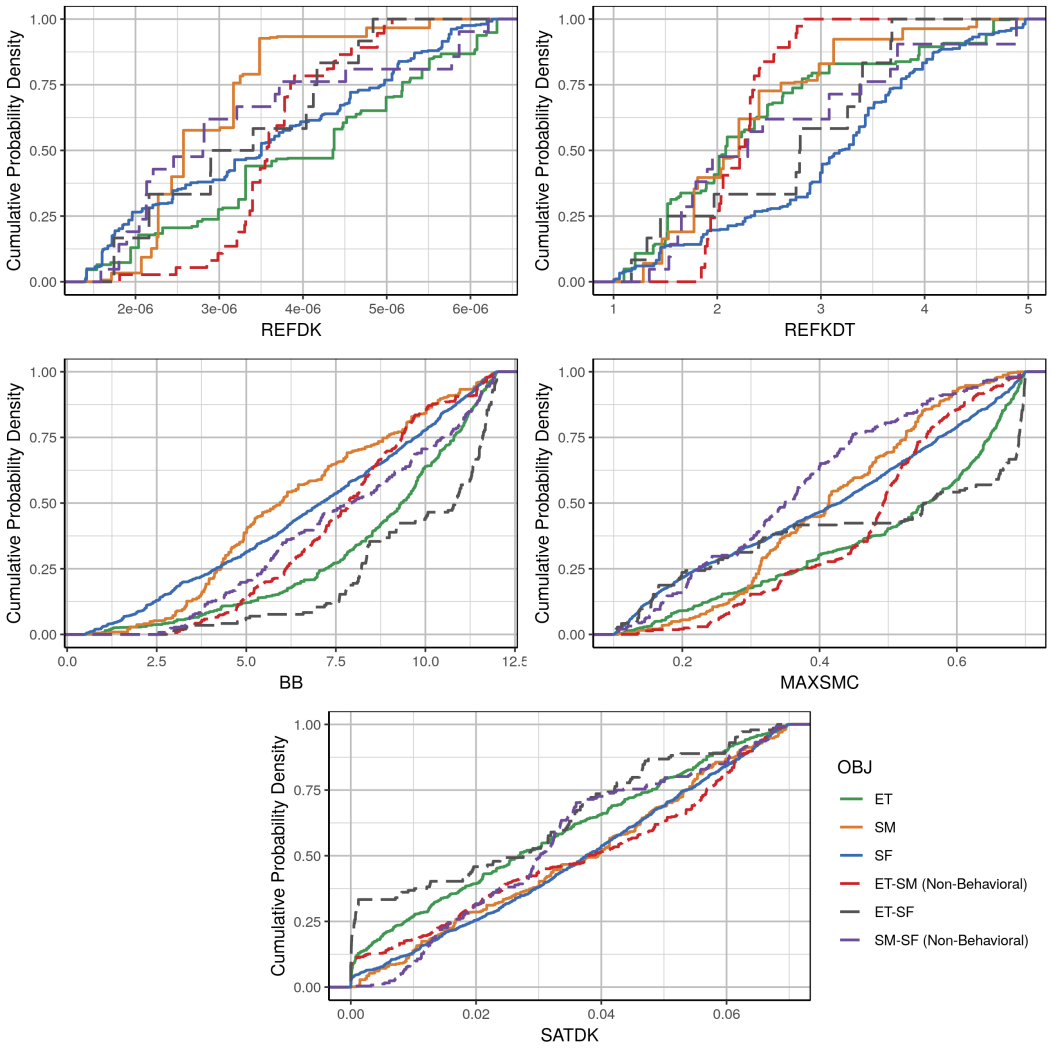

Figure 3.7: Empirical cumulative density functions (ECDF) of calibrated Noah-MP parameters for univariate (ET, SM, and SF) and multivariate (ET-SM, ET-SF, and SM-SF) objectives. For the ET, SM, SF, and ET-SF objectives, the solutions within the error threshold of the 50% quantile are used to construct the PDFs and ECDFs of the parameters. For the ET-SM and SM-SF solutions that are not behavioral for both, model responses at the 50% quantile error threshold are used.

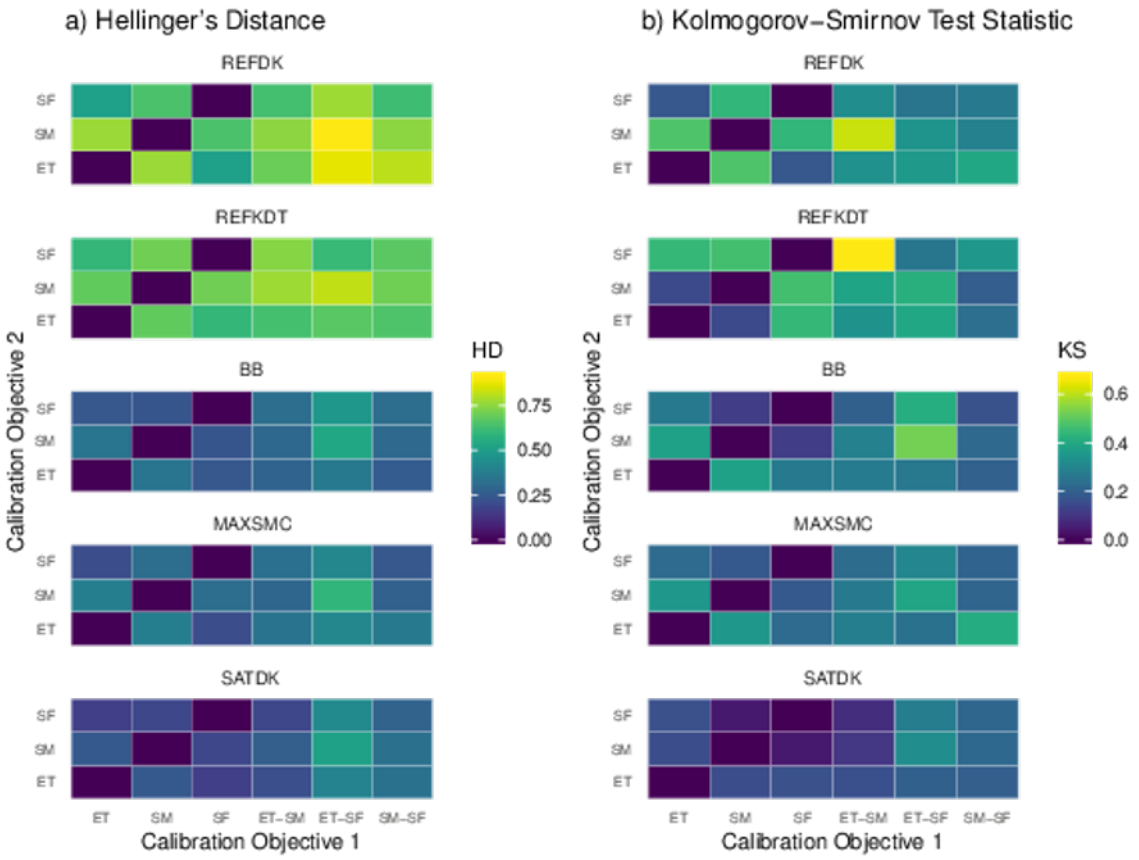

Figure 3.8: a) Hellingers distance between PDFs of the calibration parameters for different calibration objectives and b) the Kolmogorov-Smirnov test statistic between the ECDFs of calibration parameters for dierent calibration objectives. For the ET, SM, SF and ET-SF objectives, the solutions within the error threshold of the 50% quantile are used to construct the PDFs and ECDFs of the parameters. For ET-SM and SM-SF solutions that are not behavioral for both, the model responses at the 50% quantile error threshold are used.

Table 3.7: Correlation between objective functions (RMSE) and Noah-MP model parameters (REFDK, REFKDT, BB, MAXSMC and SATDK) for three multivariate calibration cases (ET-SM, ET-SF and SM-SF)

Objectives	**REFDK**	**REFKDT**	**BB**	**MAXSMC**	**SATDK**
ET and SM	(-0.57, 0.62)	(0.50, -0.16)	(-0.37, 0.27)	(-0.32, 0.17)	(0.12, -0.05)
ET and SF	(-0.27, 0.24)	(0.09, -0.02)	(-0.17, 0.17)	(0.04, -0.07)	(-0.17, 0.18)
SM and SF	(-0.15, 0.13)	(-0.02, 0.12)	(0.24, -0.28)	(-0.21, 0.20)	(0.13, -0.10)

[a] The numbers in the brackets represent correlation of the parameters with objective functions 1 (ET in first row) and 2 (SM in first row) respectively

[b] For the BB, MAXSMC and SATDK parameters the median correlation from the 12 soil classes are presented

3.5 Conclusions and future work

With the rise in availability of satellite-based measurements of hydrologic fluxes (Lettenmaier et al., 2015), multivariate calibration is widely seen as a promising solution for improving the performance and realism of large scale hydrologic models. However, most multivariate calibration studies do not formally define any acceptable error thresholds to help one conclude whether incorporation of additional fluxes into calibration improves either the performance or the realism of a hydrologic model. In addition, apriori assumptions such as complementary relationships between the different fluxes and deterministic calibration approaches hinder rigorous testing and diagnosis of hydrologic models, as called-for by Beven (2018). In this study, we develop a framework for multivariate calibration by combining Bayesian and Pareto optimality-based calibration methodologies. This framework can be used to test whether models simultaneously can simulate multiple fluxes accurately by accepting or rejecting parameter solutions based on a defined error threshold. Applying the framework to a large scale distributed hydrologic model (Noah-MP), we find that the model simulates different combinations of fluxes (ET and SM, ET and SF, and SM and SF) with

varying degrees of acceptability. While ET and SF can be simulated accurately, we find that accurately simulating either ET or SF along with SM is associated with significant trade-offs. Analyzing the trade-offs between the model responses (Table 3.6), we find that the higher trade-offs are mainly due to the fact that ET cannot be simulated accurately by calibrating the model with the other fluxes. However, calibrating the model with ET produces lower error for SF (Figure 3.5). This highlights the advantage of using a Pareto-based calibration approach, which does not assume any subjective weights in its objectives. Unlike deterministic calibration methodologies, we use parameter distributions from DREAM and AMALGAM to identify the parameters that cause significant trade-offs in accuracy between simulated fluxes. In addition to sensitive parameters that influence the behavioral simulation of model responses, we identify parameters that influence the trade-offs. For example, in the case of the Noah-MP model tested in this study, we see that the runoff parameter REFDK and the exponent in the Brooks-Corey equation (BB) influence the trade-off between ET and SM. This not only shows the advantages of framing multivariate calibration as a Pareto optimality problem but also highlights the fact that relatively insensitive parameters (such as REFDK for ET and SM) can exert a big influence on the accurate simulation of multiple fluxes.

We note that the results and conclusions we present in this study are for a specific combination of hydrologic model, input datasets, observational data and model parameters. For a different hydrologic model or calibration approach, the value of incorporating different fluxes and the relationship between them may be different. For example, the time period considered for calibration is one year and the study area is the Mississippi river basin. The calibrated parameters may not be applicable to a river basin with different physical characteristics and hydroclimatic conditions. The higher errors in the simulated fluxes (Figure 4) may be due to the chosen calibration period of one year. This is especially true for streamflow, as only 72 data points (six basins and 12 months) were used for calibration as opposed to around 60,000 data points (all the grid cells) for ET and SM. Increasing the calibration period or the number of parameters may lead to improved accuracy in the simulation of the fluxes,

and hence lower behavioral thresholds and improved Pareto optimal solutions. However, the multivariate calibration framework developed in this study is model- and data-independent and can be used to analyze the value of every flux under consideration. Future work involves extending the current study in several ways. First, we wish to incorporate more than two fluxes into the calibration to see if it changes the nature of trade-offs. For example, in the case of ET and SM, it would be interesting to see if incorporating SF reduces the overall trade-offs in accuracy among the fluxes. As REFDK is a runoff-related parameter that affects the trade-off between ET and SM, incorporating SF could lead to better discovery of REFDK. Second, incorporating estimates of total water storage (TWS) could yield better performance than the near-surface soil moisture estimates used in this study. Third, applying the developed framework to different hydrologic models and conducting inter-model comparison studies would help in model selection. Finally, the developed framework can help in the development, testing and diagnosis of new hydrologic models and model hypotheses from a multivariate perspective.

CHAPTER 4

Seasonal Hydropower Planning For Data Scarce Regions Using Multi Model Ensemble Forecasts, Remote Sensing Data, and Stochastic Programming

4.1 Introduction

Multistage stochastic programming with recourse models, widely used for the optimization of reservoir operations at seasonal time scales (Yeh, 1985), utilize scenario trees to incorporate the uncertainty in future inflows (Uysal et al., 2018). Scenario trees are generally constructed using statistics from reliable long-term streamflow records (Trezos and Yeh, 1987) or ensemble streamflow predictions (Alemu et al., 2011). In data scarce regions, such as sub-Saharan Africa which has witnessed exponential growth in installed hydropower capacity (Conway et al., 2015, 2017), scenario generation methods based on historical streamflow observations are ineffective owing to unreliable and inadequate streamflow records. At shorter lead times, streamflow forecasts generated by forcing calibrated hydrologic models with ensemble precipitation forecasts from numerical weather prediction (NWP) models have been used to construct inflow scenario trees (Lee et al., 2008; Saavedra Valeriano et al., 2010; Wang et al., 2012). At seasonal time scales, scenario trees are valuable as streamflow forecasts have large uncertainty due, mainly, to lack of skill in long-lead precipitation forecasts (Shukla and Lettenmaier, 2011). However, the development of ensemble climate prediction systems has resulted in improved forecasts of precipitation at seasonal time scales (Lavers et al., 2009). While the combination of seasonal precipitation forecasts and hydrologic models has the potential to generate inflow scenarios in data scarce regions (Block et al., 2009), it

does not completely obviate the need of streamflow observations. Streamflow measurements are required for calibrating hydrologic models. For the calibration of hydrologic models, satellite-based observations of water balance components, such as evapotranspiration (ET), soil moisture (SM), snow water equivalent (SWE), and total water storage (TWS) (Lettenmaier et al., 2015), can potentially be considered as proxies for streamflow. In this study, we propose and test a framework for seasonal hydropower planning in data scarce regions where streamflow measurements are either unreliable or unavailable. Specifically, our research focuses on 1) the development of a scenario generation methodology that combines ensemble seasonal precipitation forecasts, spatially distributed hydrologic models as well as satellite-based estimates of hydrologic variables and 2) the reformulation of the classical stochastic programming with recourse model to take into account the uncertainties in seasonal forecasts.

Reservoir inflow scenarios generated from hydrologic models can take into account three primary sources of uncertainty - 1) the uncertainty in precipitation forecasts (Shukla and Lettenmaier, 2011), 2) the uncertainty in the estimates of initial hydrologic conditions (such as snow water equivalent and soil moisture) (Koster et al., 2010; Mahanama et al., 2008) and 3) the uncertainty in model parameters (Demirel et al., 2013). Multi model ensembles (MME) are widely used for characterizing the uncertainty in precipitation forecasts, arising primarily from model physics and initial atmospheric conditions (Krishnamurti et al., 2016; Shrestha et al., 2015). For seasonal climate forecasts, a number of MMEs developed in the past decade such as DEMETER (Palmer et al., 2004), ENSEMBLES (Weisheimer et al., 2009), the North American Multi Model Ensemble (NMME) (Kirtman et al., 2014), and the NCEP Climate Forecast System (CFSv1 and CFSv2) (Saha et al., 2006, 2014) have shown promise. For example, the NCEP CFSv2 MME forecasts precipitation with reasonable accuracy at shorter leads (1-2 months) over different hydroclimatic regions (Yuan et al., 2011; Siegmund et al., 2015). Studies evaluating the forecast capability of NMME conclude that the ensemble members have a realistic spread and, in general, the ensemble mean outperforms the individual members (Becker and van den Dool, 2016; Becker et al., 2014; Cash

et al., 2017). Remotely sensed SM and SWE datasets can be used to characterize the uncertainty in the estimates of initial hydrologic conditions. Techniques such as the ensemble Kalman filter (EnKF) have found widespread use for soil moisture (Yatheendradas et al., 2012) and SWE (Alvarez-Garreton et al., 2016) data assimilation. In the absence of reliable streamflow records, the potential of remotely sensed ET and SM observations in characterizing the model parameter uncertainty needs to be explored. A number of calibration studies have quantified the value of using ET and SM datasets for informing streamflow simulations. Wanders et al. (2014) calibrated the LISFLOOD hydrologic model with remotely sensed SM from AMSR-E, SMOS, and ASCAT data. The results of the study show improved streamflow performance compared to an uncalibrated model. In a similar study that used satellite-based ET measurements for calibrating the SWAT hydrologic model, streamflow simulations improved over the base model (Immerzeel and Droogers, 2008). It has been shown that multivariate calibration using both ET and SM datasets can provide substantial improvements in the accuracy of streamflow simulations compared to single objective calibration (López López et al., 2017). In this study, we use evapotranspiration as a proxy for streamflow in calibrating the hydrologic model. We use a Bayesian calibration methodology to quantify the impact of model parameter uncertainty on the optimal release policies.

The classical formulation of a stochastic programming with recourse model considers the first stage to be deterministic and the subsequent stages to be stochastic (Yeh, 1985). In other words, it is assumed that a single deterministic inflow forecast is sufficiently accurate for the first stage. This assumption is valid when the optimization horizon is short (0-10 days) and the inflow forecast uncertainty is low, as seen in a number of studies. For example, Xu et al. (2014) consider the inflow forecast at the first stage to be sufficiently accurate in developing a Bayesian hydropower optimization model with a horizon of 5 and 10 days. In another study, precipitation forecasts from a deterministic NWP model were used to generate streamflow forecasts for the first stage and for the second stage a 30-member multi-model ensemble was used (Wang et al., 2012). Even when the planning horizon is seasonal to inter-annual, the first stage (month) inflow is assumed to be deterministic and

the optimal release is continuously updated when more observations of the current months inflow become available (Kim and Palmer, 1997). For example, Etkin et al. (2015) develop and test a stochastic decision support tool for seasonal multipurpose reservoir operation in which the immediate stage is deterministic. In data-scarce catchments where streamflow measurements are not available and the accuracy of seasonal streamflow forecasts cannot be reliably quantified, the assumption of deterministic first/immediate stage fails. Sguin et al. (2017) compare scenario trees of varying complexities 1) Full scenario tree with first stage deterministic, 2) scenario tree with only the median scenario at all stages and 3) scenario fan and conclude that, for hydropower planning, stochastic scenarios (scenario tree or fans) are preferable over deterministic scenarios. In this study, we reformulate the stochastic programming with recourse model for hydropower planning to take into account the uncertainty in seasonal reservoir inflow forecasts at the first/immediate stage. In summary, we address the following research questions: 1) Can seasonal precipitation forecasts combined with remotely sensed ET observations generate reliable reservoir inflow scenarios in the absence of streamflow observations? 2) How do uncertainties in precipitation forecasts and model parameters impact seasonal reservoir inflow forecasts, and hence hydropower production? 3) To what extent does incorporation of inflow uncertainty in the first/immediate stage of a stochastic programming with recourse model affect the release policy, and hence hydropower production?

4.2 Methodology

Consider a cascade of R reservoirs. Let the seasonal hydropower planning horizon be T months, which is divided into T stages (each stage is one month). Let the monthly time period be $t, t = 1, 2, ...T$. In a multistage stochastic programming with recourse model, at the beginning of each time period t, a reservoir inflow scenario tree or fan is constructed to represent possible future inflows into the reservoir. These scenario trees are used as an input to an optimization model that minimizes or maximizes the expected value of a specified objective function. In this study, we adopt a rolling horizon scheme, in which the optimal

release decisions are adopted only for the immediate stage (t). Then, the planning horizon is rolled forward ($t = t + 1$), the reservoir inflows are reforecast, and the hydropower is re-optimized. This process is repeated until the end of the planning horizon ($t = T$). In the classical formulation of a stochastic programming with recourse model, the scenario trees or fans are constructed in such a way that the immediate stage is deterministic. Note that t starts from 1, i.e., the beginning of the first month and moves forward with a rolling horizon of T. Due to higher uncertainty in seasonal forecasts of inflows in catchments where streamflow observations are not available, this assumption may not hold. In our study we evaluate three different scenario tree structures (Figure 4.1) and their impact on hydropower production in data scarce regions 1) a single deterministic forecast (DET), 2) a scenario fan, but the first stage is deterministic (SPWR-D) and 3) a scenario fan with all stages stochastic (SPWR-S). Irrespective of the scenario structure, we combine a spatially distributed hydrologic model with multi model ensemble (MME) precipitation forecasts from NWP models to generate reservoir inflow forecasts. We first describe the challenges in generating reliable inflow forecasts in data scarce regions. Then we detail the approach used to generate the deterministic forecast (DET) and the scenario fans (SPWR-D and SPWR-S).

To generate reliable reservoir inflow forecasts in catchments where streamflow measurements are not available, we use evapotranspiration (ET) as a proxy for streamflow (SF). The selection of ET as a proxy for SF is motivated by the findings of several previous studies which conclude that incorporating ET, in the absence of SF observations, into calibration improves the accuracy of SF compared to an uncalibrated model (Immerzeel and Droogers, 2008; López López et al., 2017; Zink et al., 2018). Several previous studies have tested the validity of combining hydrologic models and ensemble precipitation forecasts for generating inflow forecasts for reservoir optimization at different time scales. For example, Lee et al. (2008) use the rainfall-runoff forecasting system (RRFS) hydrologic model to translate deterministic and ensemble precipitation forecasts from the Korean regional and global data assimilation systems (RDAPS and GDAPS) into reservoir inflow forecasts at 1-day to 10-day lead times. In another study, a thirty-member reservoir inflow ensemble is generated

for short-term (8-day lead time) reservoir optimization by using precipitation forecasts from deterministic NWP models and perturbed quantitative precipitation forecasts (QPFs), and the WEB-DHM hydrologic model (Wang et al., 2012). Etkin et al. (2015) construct a multi stage inflow scenario tree using the ABDC rainfall-runoff model and precipitation scenarios derived from historical data for a planning horizon of one year. In all these studies, NWP model-based precipitation forecasts are used at sub-seasonal time scales. In addition, the hydrologic model is calibrated and/or validated using gauge-based measurements of streamflow. In our study, we use satellite-based ET estimates for calibrating and validating a hydrologic model, which is then used to generate reservoir inflow forecasts at seasonal time scales. In addition, to quantify the model parameter uncertainty, its impact on the inflow forecasts, and hence the hydropower, we adopt a formal Bayesian calibration approach to derive the posterior probability distribution of model parameters. Specifically, we utilize the Differential Evolution Adaptive Metropolis (DREAM) Markov Chain Monte Carlo Scheme (MCMC) scheme (Vrugt et al., 2009a, 2008), which has been extensively used to quantify parameter uncertainty in hydrologic (Shafii et al., 2014) and hydrogeologic models (Laloy et al., 2013). First, the hydrologic model is calibrated using the DREAM algorithm to derive the posterior distribution of model errors. Next, we define a limit of acceptability or error threshold to distinguish between behavioral and non-behavioral solutions (Beven, 2006). Behavioral solutions are model parameter sets, derived from calibration, that result in errors that are within an acceptable limit (error threshold) for a specific model output (ET in this study). We use the model parameter set with the least model error for generating inflow scenarios. The remaining behavioral model parameter sets are then used to quantify the uncertainty in ET and SF forecasts due to uncertainty in model parameters. We describe the setup, calibration and validation of the hydrologic model used in our study, and the configuration of the DREAM calibration algorithm in the next (experiment design) section. To generate the deterministic forecasts in the DET and SPWR-D scenarios (Figure 4.1), we use Bayesian Model Averaging (BMA) to calibrate ensemble precipitation forecasts. Then, the calibrated deterministic precipitation forecast is used an input into the calibrated hydrologic model to generate the deterministic inflow forecast. BMA is a statistical approach to post processing

forecast ensembles generated from multiple statistical (Hoeting et al., 1999) or dynamical models (Raftery et al., 2005). BMA of a forecast ensembles results in a calibrated and sharp predictive probability density function (PDFs), represented as a weighted averaged of the PDFs of the ensemble members. Following Sloughter et al. (2007), the BMA predictive PDF can be mathematically represented as

$$P(x|f_1, f_2, ..., f_n) = \sum_{n=0}^{N} \omega_n c_n(x|f_n) \tag{4.1}$$

where ω_n is the posterior probability of ensemble forecast member f_n being the best one, determined in the calibration or training period using observed or reference data of the hydrologic variable under consideration x (for example, precipitation or evapotranspiration), $c_n(x|f_n)$ is the conditional PDF associated with the ensemble forecast f_n of the hydrologic quantity x. For variables such as temperature, the conditional PDF can assumed to be normally distributed (Raftery et al., 2005), but a gamma distribution is more appropriate for precipitation (Sloughter et al., 2007), evapotranspiration (Khanmohammadi et al., 2018), and streamflow (Vogel and Wilson, 1996). In our study, we use a mixture of point mass at zero and a gamma distribution as the conditional PDF (Sloughter et al., 2007). We determine the BMA weights (ω_n) of ensemble precipitation members using satellite-based estimates of precipitation. To calibrate the ensemble precipitation forecasts (determine the BMA weights), we define a calibration period which precedes the hydropower planning horizon. The length of the calibration period is equal to the length of the planning horizon. In other words if the planning horizon starts from month t and ends at month T, the BMA calibration period starts from the month $t - T$ and ends at month $t - 1$. We adopt a rolling scheme similar to the optimization model, wherein we recalculate the BMA weights at every stage of the stochastic programming with recourse model, by moving the calibration period forward by one month. The deterministic forecasts for the DET and SPWR-D scenario structures are then generated as the weighted average of the ensemble precipitation forecast members under consideration. The deterministic forecasts are then used as an input into the hydrologic model to generate the deterministic reservoir inflow forecast. Figure 4.2 shows

the flow charts of the proposed methodology for seasonal hydropower planning in data scarce regions using the DET, SPWR-D and SPWR-S reservoir inflow scenarios.

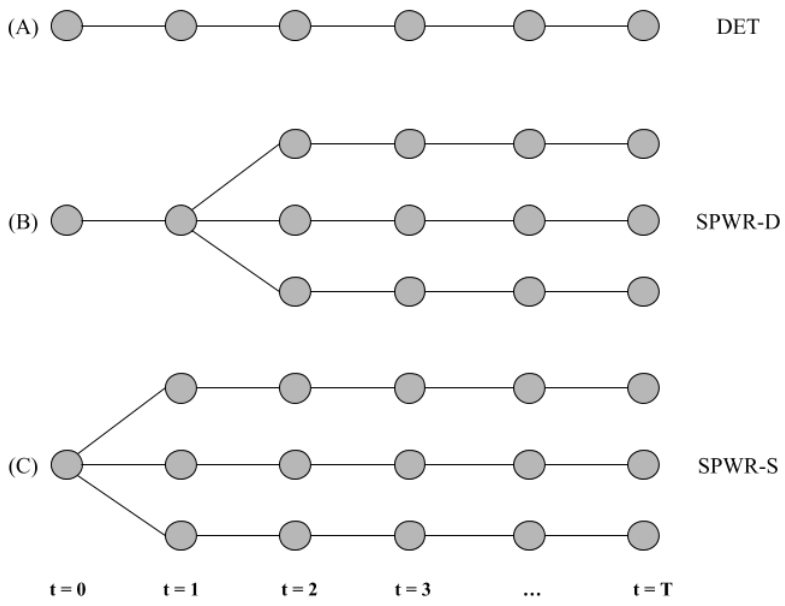

Figure 4.1: A visual representation of three scenario fan structures A) A single deterministic forecast (DET), B) first stage deterministic and the rest stochastic (SPWR-D) and C) all stages stochastic (SPWR-S).

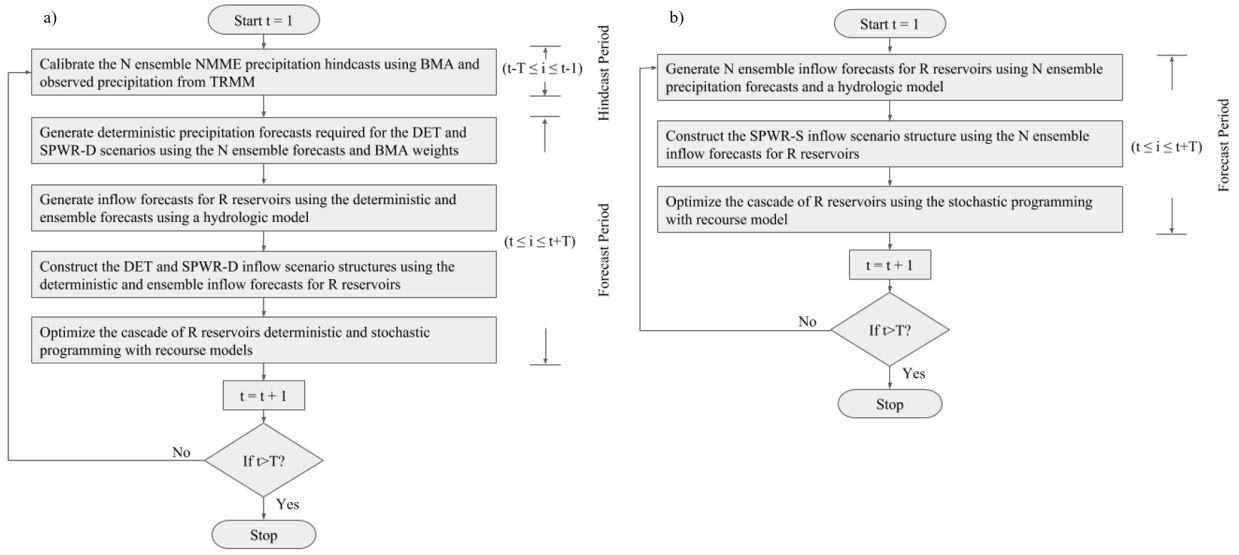

Figure 4.2: Flow chart for seasonal hydropower planning using a) DET and SPWR-D and b) SPWR-S reservoir inflow scenario structures. The precipitation forecasts, observational datasets, hydrologic model, and optimization algorithm used in the case study are mentioned in parenthesis.

4.3 Experiment design

4.3.1 Study area and time period

The seasonal hydropower planning framework developed in this study is tested in the Omo-Gibe river basin in East Africa (Figure 4.3). Spanning an area of 79,000 km^2, the river basin is spread across countries of Ethiopia and Kenya, and the outlet of the basin is at Lake Turkana. The elevation ranges from about 700 m to 3,100 m. The average rainfall in the basin is about 1,150 mm with a humid north and arid south. The annual temperature varies between 17 ^{o}C and 29 ^{o}C (Chaemiso et al., 2016). In terms of hydropower capacity, the study area consists of three power plants in operation (Gilgel Gibe I, II and III) and two planned hydropower plants (Gilgel Gibe IV and V). Table 4.1 presents the details of the hydropower plants. To test the proposed framework we use a cascade of two reservoirs, consisting of Gibe I and Gibe III. Inflows to the reservoirs are forecasted eight months in

advance, and the planning horizon is assumed to be eight months. The system is optimized with an eight-month rolling horizon. To test the framework, we select February 2005 to September 2005 as the planning horizon. We calibrate the hydrologic model for the year 2004 using satellite-based estimates of ET.

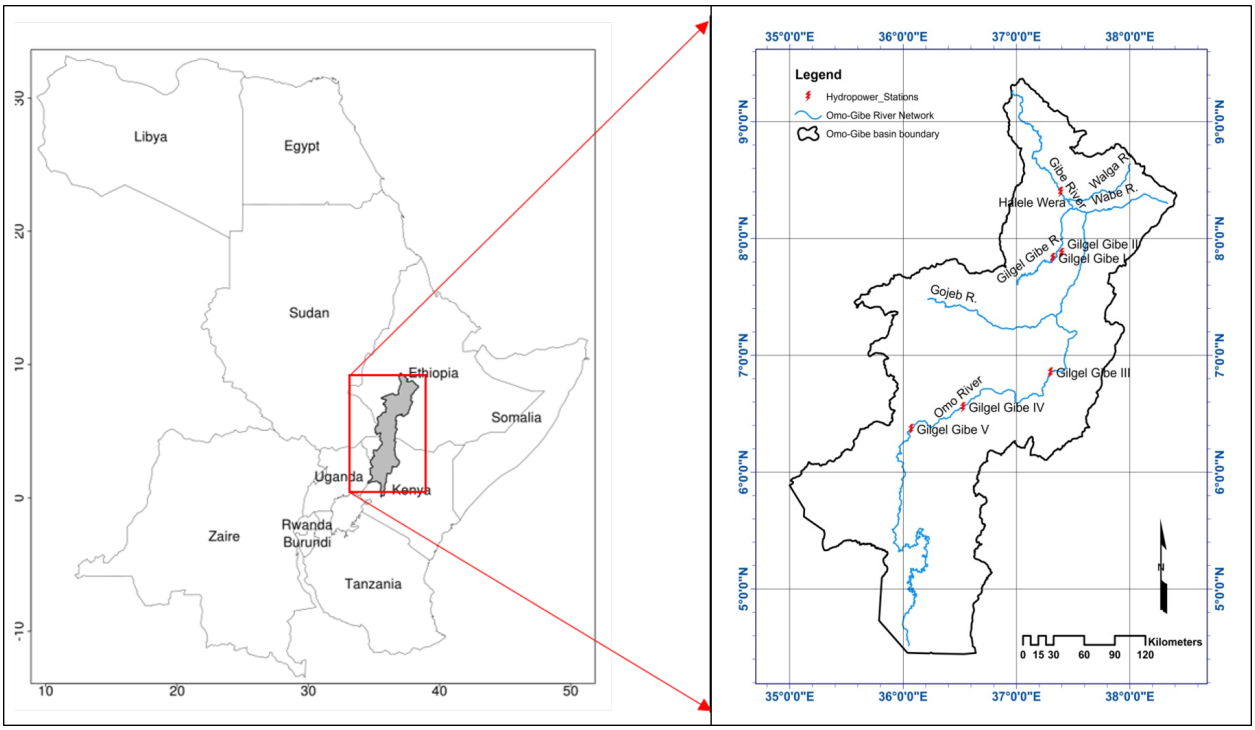

Figure 4.3: The Omo-Gibe river basin consisting of a cascade of five reservoirs, located in East Africa. The countries presented constitute the East African Power Pool (EAPP).

4.3.2 Observational and forecast data

In the study area (Omo-Gibe river basin), we use monthly estimates of ET from the Global Land Evaporation Amsterdam Model (GLEAM) (Martens et al., 2016) as observational data for calibrating the hydrologic model. Specifically, we use the GLEAM v3 ET dataset which assimilates remotely sensed soil moisture and vegetation optical depth from multiple satellites. The spatial resolution of the GLEAM dataset is 0.25^o x 0.25^o. Similarly, the precipitation input for calibrating the hydrologic model in the Omo-Gibe basin is the Tropical Rainfall Measuring Mission (TRMM) Multi-satellite Precipitation Analysis (TMPA) (Huff-

Table 4.1: Details of the five hydropower reservoirs in the Omo-Gibe river basin

Reservoir Name	**Start of Operation**	**Hydropower Capacity**	**Maximum Storage**
Gilgel Gibe I	2004	210 MW	840 Mm^3
Gilgel Gibe II	2010	420 MW	0.15 Mm^3 (Run-of-river)
Gilgel Gibe III	2015	1870 MW	13700 Mm^3
Gilgel Gibe IV	Planned	1450 MW	-
Gilgel Gibe V	Planned	600 MW	-

man et al., 2007). Specifically, we utilize the real time version (TMPA 3B42RT). The spatial resolution of TMPA 3B42RT dataset is 0.25^o x 0.25^o and the temporal resolution is three-hourly. We select GLEAM ET based on the findings of Koppa and Gebremichael (2017) in which multiple satellite-based ET and precipitation datasets were ranked using a framework based on the Budyko hypothesis (Budyko, 1974).

For generating the seasonal reservoir inflow forecasts, we use the ensemble seasonal precipitation forecasts from the North American Multi-model Ensemble (NMME) (**?**). The NMME consists of nine partner models with the number of ensemble members in each model varying from six to twenty eight. Out of the nine models we select the following three models: 1) Goddard Earth Observation System version 5 (GEOS-5) (Borovikov et al., 2017), 2) Third generation Canadian Coupled Global Climate Model (CanCM3) (Merryfield et al., 2013), and 2) Fourth generation Canadian Coupled Global Climate Model (CanCM4) (Merryfield et al., 2013). With ten ensemble members for each model, our study uses a total of thirty ensembles from three models. The hindcasts of NMME are available for the time period 1981-2010. The spatial resolution is 1.0^o x 1.0^o and the temporal resolution is daily.

4.3.3 Hydrologic model: setup, calibration and model parameter uncertainty

To translate the NMME precipitation forecasts into reservoir inflow forecasts, we choose the Noah-MP (Multi-Parameterization) Land Surface Model (LSM) (Niu et al., 2011), driven through NASAs Land Information System (LIS) (Kumar et al., 2006). The Noah-MP model builds on the original Noah LSM by incorporating a dynamic groundwater model, improved representation of vegetation canopy and snow pack. All the static input datasets required for running the Noah-MP model are sourced from NASAs LIS data portal. The important static input datasets are the land cover map, sourced from USGS; the soil texture map from STATSGO, sourced from USDA; and the elevation map from GTOPO30, sourced from USGS. Albedo, greenness fraction and temperature are sourced from NCEP reanalysis. The meteorological forcings required by the Noah-MP model include precipitation, air temperature, surface pressure, specific humidity, wind speed, and radiation. Barring TRMM precipitation, all meteorological forcings are derived from the Global Data Assimilation System (GDAS), sourced from the Environmental Modeling Center (EMC) of the National Center for Environment Protection (NCEP) (Derber et al., 1991). The spatial resolution of the dataset is 0.47^o x 0.47^o. The Noah-MP model is set up at a spatial resolution of 5km x 5km. The meteorological inputs, including the forecasts, are interpolated onto the model grid using bilinear interpolation. The model is spun-up for a period of 68 years by looping through the year 2003 until the groundwater and soil moisture storage reach equilibrium. The model time step is three hours. The number of soil layers in the model is four with thicknesses 10cm, 30cm, 60cm, and 100cm. Specific Noah-MP model physics options selected for different processes are detailed in Table 3.1.

The Noah-MP model contains 71 standard parameters (present in user-defined tables) and 139 hard-coded parameters (present in the model code). The Noah-MP model output has been found to be sensitive to about two-thirds of the 71 standard parameters (Cuntz et al., 2016). As the study is a calibration experiment involving multiple calibration cases, we keep the dimension of the calibration problem manageable by selecting five of the most sensitive parameters from the Cuntz et al. (2016) study. The selected parameters are two surface runoff

related parameters (REFDK and REFKDT), the exponent in the Brooks-Corey equation (BB), soil porosity (MAXSMC), and hydraulic conductivity at saturation (SATDK). Of the five parameters, BB, MAXSMC, and SATDK are related to soil texture. As there are twelve soil texture classes, the total number of parameters selected for calibration of the Noah-MP hydrologic model is 38 (Table 3 presents a detailed breakdown of the parameters with maximum and minimum values used for calibration). The minimum and maximum values of the parameter ranges are selected from literature (MAXSMC and SATDK ranges from Cai et al. (2014), BB and REFDK ranges from Cosby et al. (1984), and REFKDT range from Mendoza et al. (2015)). The minimum and maximum values are adjusted to improve the rate of convergence of the calibration algorithms.

To calibrate the hydrologic model and quantify the uncertainty in reservoir inflow forecasts due to uncertainty in model parameters, we use the DREAM algorithm. DREAM is a multi-chain Markov chain Monte Carlo (MCMC) simulation algorithm that automatically tunes the scale and orientation of the proposal distribution en route to the target distribution. It is designed for increasing the sampling efficiency of complex, high-dimensional parameter spaces, while maintaining detailed balance and ergodicity (Vrugt, 2016). In this study, we use the MT-DREAM (ZS) version of DREAM which utilizes multi-try sampling (MT), snooker updating and sampling from an archive of past states to improve the rate of convergence and can make use of parallel computing resources. Specific configuration options and parameters of the MT-DREAM (ZS) algorithm used in this study are detailed in Table 4. We select the Laplacian likelihood based on the findings of Schoups and Vrugt (2010); residual errors in rainfall-runoff models are better represented by a Laplacian distribution than a Gaussian distribution. The likelihood function is used to summarize the distance between the model simulations and the corresponding observations. In the absence of streamflow observations, we use satellite-based estimates of ET to calibrate the hydrologic model for the year 2004. Specifically, error residuals determined at all of the 5km x 5km grid cells and time steps (monthly) over the entire Omo-Gibe river basin. On a workstation with 16 processors, MT-DREAM (ZS) required around 12,000 iterations to converge to a

solution.

4.3.4 Hydropower optimization model

In this study, the formulation of the hydropower optimization model is based on HIDROTERM, a nonlinear programming optimization model previously developed for planning the operation of the Brazilian hydrothermal system (Zambon et al., 2012a). The model, originally deterministic, was modified to solve the multi-stage stochastic programming with recourse model for our study (Zambon et al., 2012b). The objective function is represented by:

$$minZH = \sum_{s} \sum_{t} \{p_S dt_t (D_t - \sum_{i} P_{i,s,t})^2\} \tag{4.2}$$

where: i = hydropower plant/reservoir index; dt_t = time period duration (106 s); s = scenario index; t = time period index; p_s = probability associated with each scenario; $P_{i,s,t}$ = power production (MW); D_t = objective demand, usually the total demand minus the fixed generation, though it can be defined arbitrarily by the user (for example, as the maximum installed power capacity MW); ZH = model objective (106 s.MW^2).

The model minimizes the expected value of the quadratic departures from the demand so that the hydropower production will follow the specified demand variations. The model is subject to the following additional set of constraints: total release is a summation of the power release (turbine) and non-power release (spill); continuity equation for the storage reservoirs, including evaporation losses; bounds for storage, releases and power production; water head as difference from reservoir forebay and tailrace water levels; power production as function of water head and power release; reservoir level-area-storage curves. The deterministic equivalent of the stochastic model is solved by nonlinear programming (NLP) using the General Algebraic Modeling System package (GAMS, 2018) and considers individual hydropower plants. The deterministic scenario (DET) has 32 decision variables and 320 constraints. The SPWR-D scenario has 844 decision variables and 8720 constraints. The SPWR-S scenario has 960 decision variables and 9600 constraints. The execution times of

each optimization run for DET, SPWR-D, and SPWR-S scenarios are on an average 0.28 seconds, 1.65 seconds, and 2.01 seconds respectively.

4.4 Results and discussion

First, we validate the seasonal precipitation hindcasts from NMME using the satellite-based TRMM dataset. Next, we validate the calibrated Noah-MP hydrologic model with remote sensing- based ET estimates from GLEAM. We then present the seasonal hydropower planning results derived from HIDROTERM. Finally, we quantify the uncertainty in optimal reservoir power release decisions and hydropower arising due to uncertainty in model parameters.

4.4.1 Validation of the NMME precipitation forecasts

To validate the 30-member seasonal ensemble precipitation forecasts from NMME, we use the TRMM remotely sensed precipitation estimates. Figure 4.4 presents a time series comparison of different precipitation forecast models with the observed data. We also present Taylor diagrams (Taylor, 2001) to represent the root mean square error (RMSE), correlation with observations and standard deviation of different forecast models (Figure 4.5). We see that the thirty ensemble members, represented by the 5% and 95% quantiles (grey lines in Figure 4.4), encompasses the observed data (dashed black lines in Figure 4.4) up to lead times of three months (lead time 1-3). The months of March, May, October, and November of the year 2005 (validation time period) are the exceptions. It is seen that all the raw ensemble members underestimate the precipitation for the month of May 2005 and over estimate precipitation for the months of October and November of the same year. In addition, we see that the difference between 5% and 95% quantiles increases with increase in lead time. This is expected as the uncertainty in the initial and boundary conditions driving the NWP models is higher at longer lead times. At lead times longer than 4 months (lead times 4-8), the raw ensemble members considerably over estimate the precipitation observed in the

second half of the year 2005. To understand the differences in performance among the three NMME models under consideration (CanCM3, CanCM4, and GEOS-5), we compare the average of the ten constituent members. The stochastic programming with recourse model implemented in this study uses forecasts that are updated at the start of each month and the release decisions are implemented only for the immediate stage. Therefore, we focus on the performance of the forecasts at shorter lead times. At lead times of 1-3 months, we see that the CanCM4 model consistently outperforms the other two models (Figure 4.4 and 4.5). Both the CanCM3 and GEOS-5 models overestimate the precipitation in the second half of 2005 (validation time period). This is especially true for the lead time 1-month where the CanCM4 model exhibits lower RMSE and higher correlation with the observed data (Figure 4.5). Additionally, at lead time1, the CanCM4 model is able to capture the two peaks in the observed precipitation of the year 2005. Between CanCM3 and GEOS-5, GEOS-5 significantly underestimates the peaks in 2005. At longer lead times, CanCM4 model performs relatively poorly compared to the other models. As expected the forecast accuracy of all the models deteriorates with increasing lead times.

Finally, we compare the simple mean and Bayesian model average of all the 30 ensemble members with the individual models. We see that the ensemble BMA outperforms all the other models at almost every lead time (Figure 4.4 and 4.5). At shorter lead times (1-3 months), the ensemble BMA is able to capture the quantity and the timing of the peaks and troughs seen in the observed precipitation (Figure 4.4). The low RMSE (approximately 25 mm/month) and high correlation coefficient (greater than 0.9) at lead time 1-month supports this conclusion. At shorter lead times, the CanCM4 model outperforms the ensemble mean in terms of both the magnitude of error and the correlation with the observed data. This highlights the advantage of the Bayesian model averaging technique, which provides higher weights to better performing models (CanCM4 in this case). The ensemble mean which weighs all the ensemble members equally is biased by the lower performing CanCM3 and GEOS-5 models members. The improvements are particularly greater in the second half of the year 2005, in which the overestimation of precipitation by CanCM3 and GEOS-5

are reduced by ensemble BMA and the ensemble mean. In addition, the variance of the ensemble BMA seems to also match the variance of the observed precipitation better than the individual members, except at lead time 1-month, where CanCM4 standard deviation matches the observed value better.

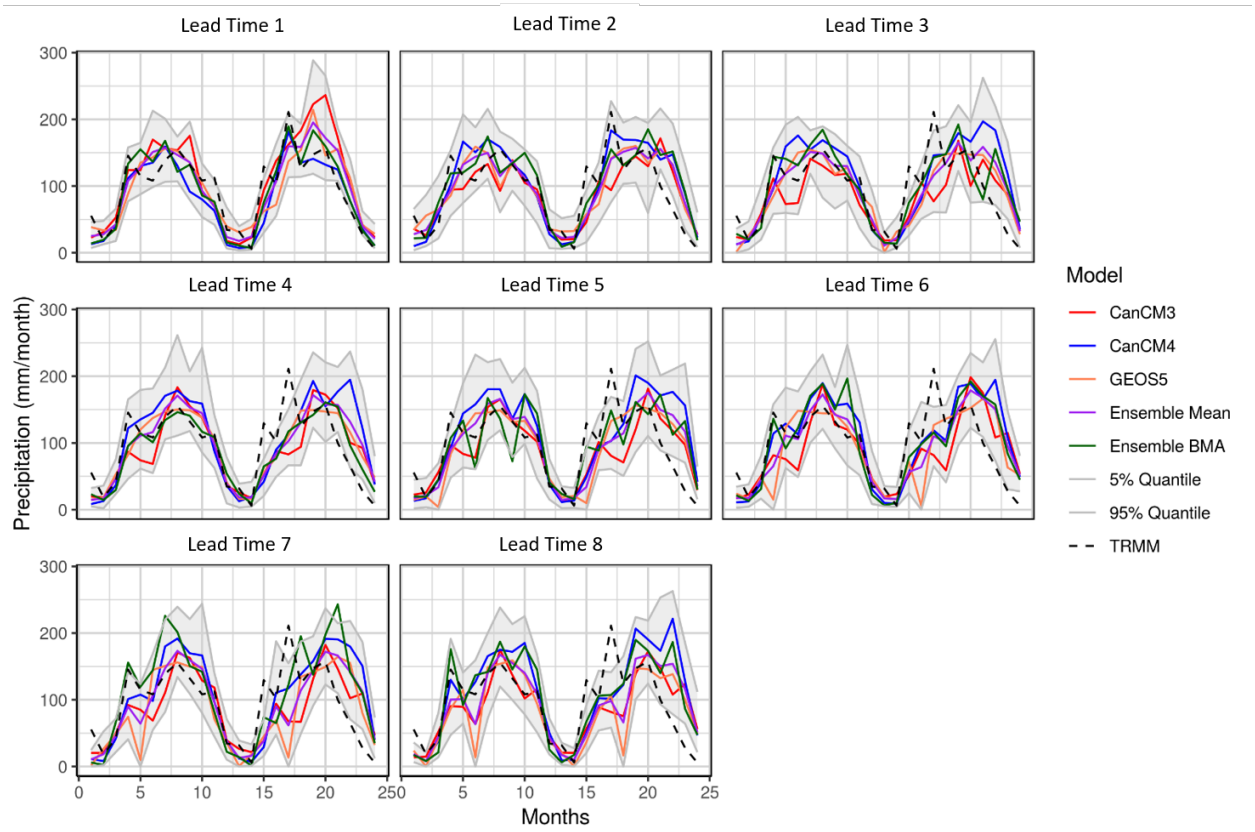

Figure 4.4: Time series comparison of precipitation from different NMME models and observations (TRMM) for different lead times (1-8 months) and for the calibration (12 months of 2004) and validation (12 months of 2005) time periods. We present the mean of the 10 members from CanCM3, CanCM4 and GEOS-5. Ensemble Mean and Ensemble BMA are the simple mean and Bayesian model average of all the 30 ensemble members. In addition, we present the 5% and 95% quantiles of all the 30 ensemble members.

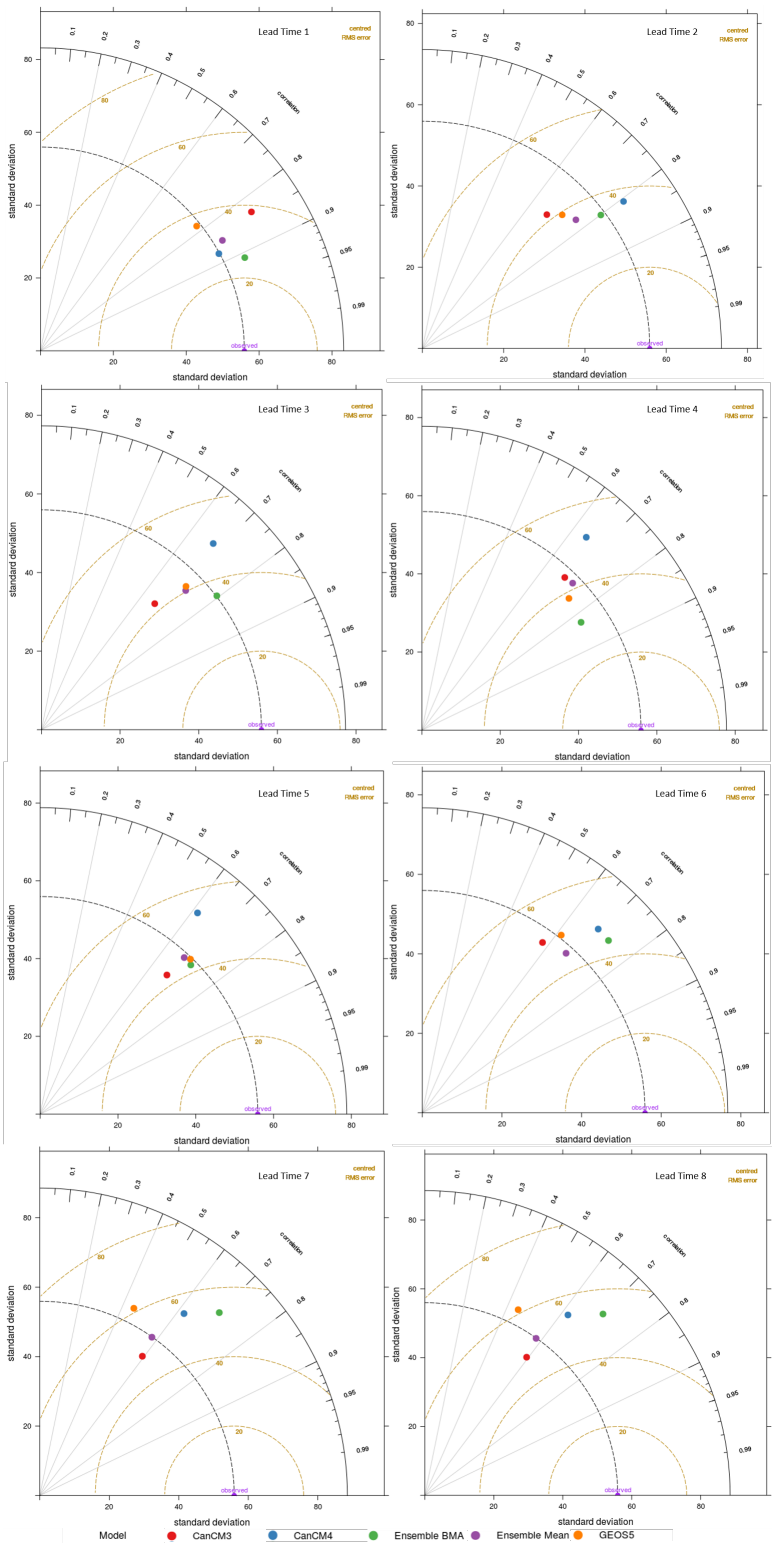

Figure 4.5: Taylor diagrams of precipitation from CanCM3, CanCM4, GEOS-5, Ensemble Mean and BMA models determined for different lead times (1-8 months) and for the calibration (12 months of 2004) and validation (12 months of 2005) time periods.

4.4.2 Validation of the Noah-MP hydrologic model

As stated in the experiment design section, we use evapotranspiration (ET) as a proxy for streamflow (SF) in calibrating the Noah-MP hydrologic model for the Omo-Gibe river basin. Accordingly, we validate the hydrologic model with remote sensing-based ET estimates from GLEAM. Figure 4.7 presents the posterior distribution of Noah-MP model parameters derived from the MT-DREAM (ZS) algorithm. We see that the distributions of all of the parameters, except REFDK, are well defined (non-uniform). The distributions of BB and MAXSMC, which are soil related parameters, are skewed towards the higher values. We can identify the high values of BB (greater than 10.0) and MAXSMC (greater thank 0.4) parameters combined with lower SATDK seen in Figure 4.7 with sandy clay, silty clay, and clay soil textures. Compared to the soil-related parameters, the distributions of the runoff-related parameters are less well defined. While the REFKDT parameter is more skewed towards the lower values, the REFDK parameter is more uniform. REFKDT is a parameter that controls the partitioning of total runoff into surface and subsurface components. A lower value of REFKDT implies that a larger portion of the total runoff is being partitioned into surface runoff. The optimized parameter set from Bayesian calibration is used to construct the posterior distribution of simulated ET errors. Then, we define an error threshold of 50% quantile to select a set of behavioral (acceptable) parameter sets. Twenty parameter sets from the behavioral solutions are used to quantify the uncertainty in inflow forecasts, and hence the hydropower, due to uncertainty in model parameters.

We validate the set of selected behavioral solutions with GLEAM ET reference dataset for the year 2005 using time series analysis (Figure 4.6) and relevant error metrics. A visual examination of Figure 4.6 reveals that the ET-calibrated Noah-MP model performs well in simulating the evapotranspiration over the Omo-Gibe river basin. In the calibration period (12 months of 2004), we see that the reference ET data (black line) is closer to the 95% quantile of the modelled ET. This implies that the behavioral solutions consistently underestimate ET. In the validation period (12 months of 2005), the reference dataset corresponds to the median modelled ET values for the first six months. In this last six months of 2005,

we see that the model consistently underestimates (except November 2005) the evapotranspiration over the study area. This consistent underestimation of ET coupled with the fact that the NMME forecasts overestimate precipitation in the second half of 2005 can lead to overestimation of reservoir inflows. It is interesting to note that the model results show higher uncertainty for the month of October in both the calibration and validation time periods compared to other months. Additionally, the model is able to capture the quantity and timing of the peaks and troughs seen in the reference ET dataset. A low mean RMSE of 3 mm/month and a high mean correlation co-efficient of 0.97 supports this conclusion. We note that the primary assumption of our study is that ET can be used as a proxy for streamflow in data-scarce regions. Consequently, we assume that the high fidelity of the Noah-MP model in simulating ET over the Omo-Gibe river basin leads to accurate simulation of streamflow, and hence the reservoir inflows required for seasonal hydropower planning.

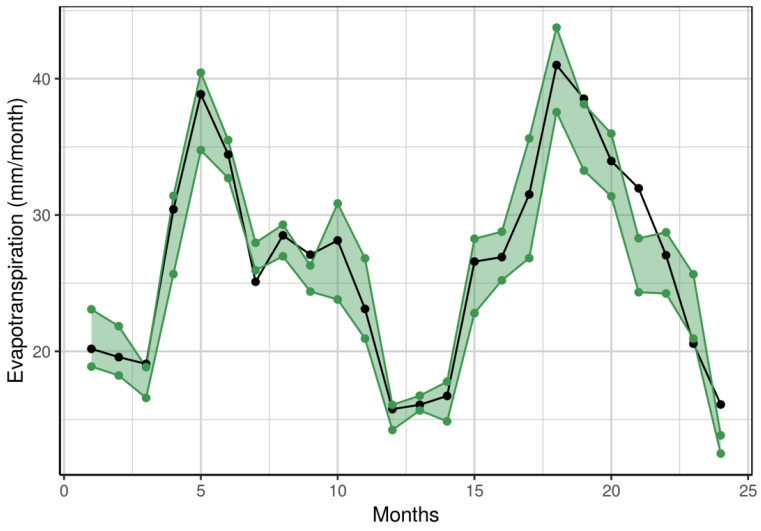

Figure 4.6: Time series comparison of evapotranspiration from ET-calibrated Noah-MP model (green) and observed ET estimates (black) from GLEAM for the calibration (12 months of 2004) and validation (12 months of 2005) time periods. The 5% and 95% quantiles from the behavioral solutions of Bayesian calibration is used to determine the uncertainty in modeled ET (green band).

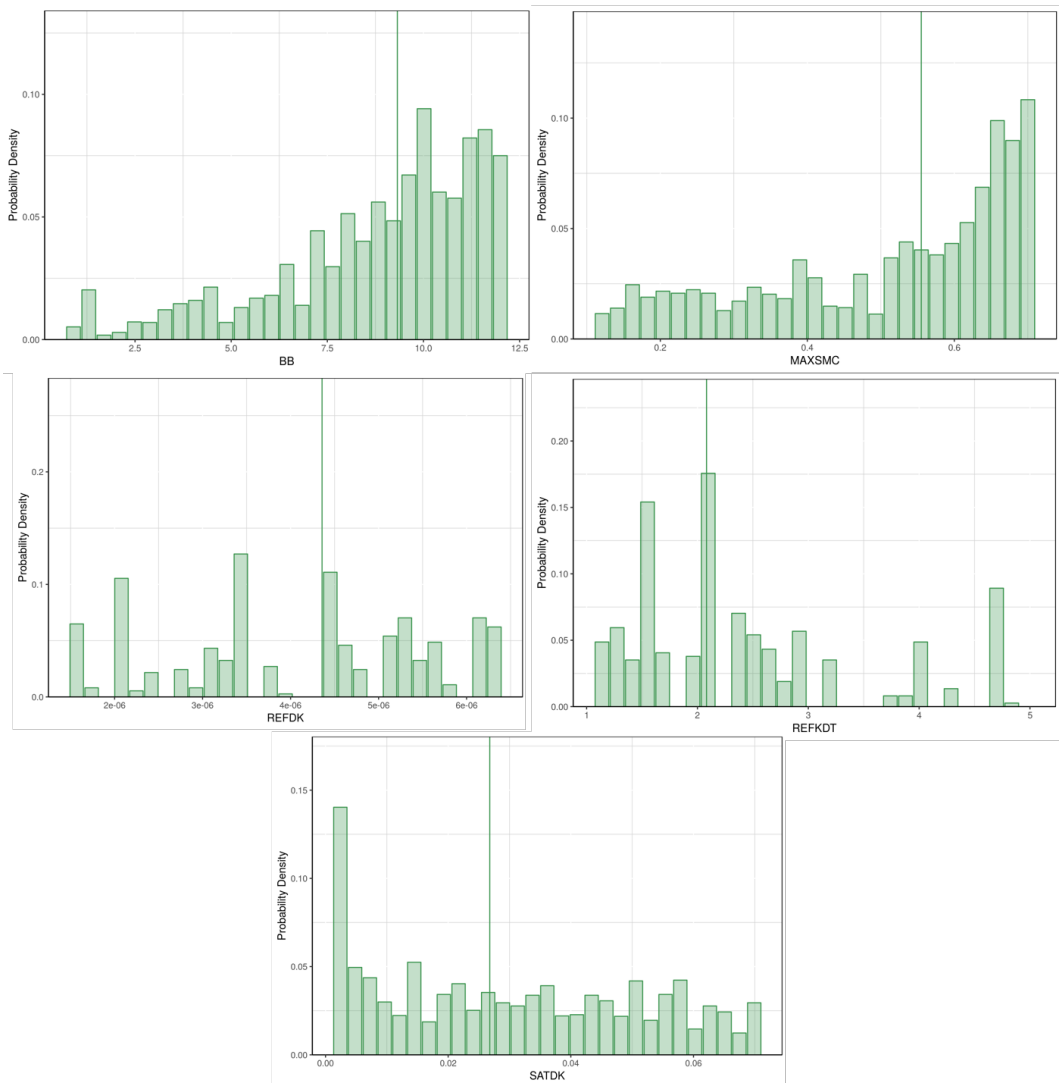

Figure 4.7: Posterior probability density functions (PDFs) of the Noah-MP hydrologic model parameters considered for calibration using the MT-DREAM (ZS) algorithm and ET estimates from GLEAM. The green line represents the 50% quantile values for each of the parameters.

4.4.3 Seasonal Hydropower planning in the study region

We select the best performing Noah-MP model parameter set from the twenty behavioral solutions to generate the reservoir inflow forecasts required to optimize the cascade of two reservoirs (Gibe I and Gibe III) in the study region. We construct the three scenario struc-

tures (DET, SPWR-D, and SPWR-S) for the months February 2005 to September 2005 by combining the NMME precipitation forecasts and the Noah-MP hydrologic model as detailed in the methodology section. The three inflow scenario structures are then used as inputs to a deterministic and stochastic programming with recourse model to generate the optimal power releases, reservoir storage variations, and the associated hydropower production (Figure 4.8). In a stochastic programming with recourse model, only the immediate stage release decisions are implemented. At the end of the immediate stage, scenario tree or fan structures are re-generated and the system is re-optimized with a rolling horizon. Figure 4.8 shows the results obtained from the optimization model for the immediate stage. First, we compare the differences in reservoir inflows among the three scenarios. We then, analyze the impact of these differences in the scenario structure on the optimized release decisions, storage variations and hydropower production. Figure 4.8a presents the deterministic (from BMA) and stochastic (raw forecast ensembles) inflows used to construct the three scenario trees. We see that the first three months of the study period are relatively dry compared to the rest of the months. Also, the deterministic inflows into both Gibe I and Gibe III are consistently lower than the mean of the 30-member ensemble, with June and July being the exceptions. Additionally, we also see that there is considerable uncertainty in the reservoir inflow arising from uncertainty in the input precipitation ensembles. To compare the uncertainty of inflows across different months and reservoirs, we calculate coefficient of variation and coefficient of range values (Table 4.2). The high values of measures of absolute and relative dispersion reflects the high uncertainty in the inflow values. The inflows in the month of March 2005 exhibit the largest uncertainty with a coefficient of variation of 0.74 and 0.69 for Gibe I and Gibe III reservoirs respectively. The month of July 2005 has the least uncertainty with a coefficient of variation of 0.31 and 0.24 for the Gibe I and Gibe III inflows respectively.

In Figure 4.8b, we present the results of the optimized release decisions generated for the three inflow scenario structures 1) DET, 2) SPWR-D, and 3) SPWR-S. We see that the optimized release decisions generated with the DET and SPWR-D inflow scenario structures

are very similar to each other, except for the month of May 2005. We note here that the results presented correspond to the immediate stage of the stochastic programming with recourse model. Therefore, the similarity between DET and SPWR-D release decisions may be due to the fact that the immediate stage of the SPWR-D model is deterministic, and the value derived from BMA of ensemble members is equal to the DET inflow scenario structure. We see that the optimized release decisions are consistently higher in the SPWR-S case compared to either the DET or SPWR-D scenario structures. This can be attributed to the fact that the inflows in the DET scenarios are consistently lower than the mean of the ensemble members (Figure 4.8a). It is interesting to note that the uncertainty in the inflows represented in the SPWR-S scenarios only impacts the dry periods (February 2005 to May 2005). In the wet months, the power releases reached the capacity of the power plants. For reference we also present the storage variations corresponding to the optimized release decisions in Figure 4.8c. Figure 4.8d presents the optimized hydropower productions corresponding to the optimized release decisions for each of the three inflow scenario structures. The differences among the three inflow structures match the differences in optimal release decisions: 1) In the dry periods (Feb 2005 to May 2005), the SPWR-S inflow scenario structure leads to higher hydropower production compared to the DET and SPWR-D cases, 2) The differences among the scenario structures and the uncertainty in reservoir inflows do not have any impact in the wet months of the study period. Finally, we compare the optimized hydropower values with the actual power produced at the Gibe I reservoir for the months of February 2005 to September 2005. The actual power produced for the 8 month study period is 746 MW. In comparison, the optimized hydropower generated from DET, SPWR-D, and SPWR-S scenario structures are 1,014 MW, 994 MW and 1,060 MW. We note that reservoir evaporation has been ignored in our study, which, when included, may reduce the optimized hydropower values. Nevertheless, with inflow forecasts and optimization, there is a substantial gain in power production when compared with historical operational records. The higher power produced in the SPWR-S scenario structure may be attributed to the consistently higher values of reservoir inflow compared to the DET and SPWR-D scenarios.

In addition to incorporating the uncertainty in reservoir inflow forecasts using stochastic programming, we quantify the impact of model parameter uncertainty on inflow forecasts, power release decisions, and hydropower production (Figure 4.9). Comparing Figure 4.8a and 4.9a, we see that the range and standard deviation of reservoir inflows have increased for all the months in the study period. We present the specific values of range and standard deviation values in Table 4.3. However, we see little difference in the coefficient of variation and coefficient of range values, with the exception of April 2005. This indicates that model parameter uncertainty does not have a significant impact on the deterministic and stochastic inflow forecasts. The corresponding release decisions (Figure 4.9b), storage variations (Figure 4.9c), and the optimized hydropower values (Figure 4.9d). With the exception of March 2005 for DET and SPWR-D scenarios, the uncertainty in the release decisions and the corresponding hydropower production due to model parameter uncertainty is not very high. It is seen that the impact of model parameter uncertainty on release decisions is more pronounced during the dry months. In the wet months, model parameter uncertainty has no impact on the release decisions. Finally, we compare the uncertainty in total hydropower production by the two reservoirs for the three inflow scenario structures. For the DET scenario, the combined hydropower production from Gibe I and Gibe III varies from 3,726 MW to 3,942 MW. Similarly, for the SPWR-D scenario structure, the combined hydropower production varies from 3,677 MW to 3,820 MW, and from 3,897 MW to 4,048 MW for the SPWR-S scenario structure. This shows that model parameter uncertainty can impact the hydropower production by approximately 200 250 MW or 5-6% of the total power, which could be significant in seasonal planning of large hydropower systems.

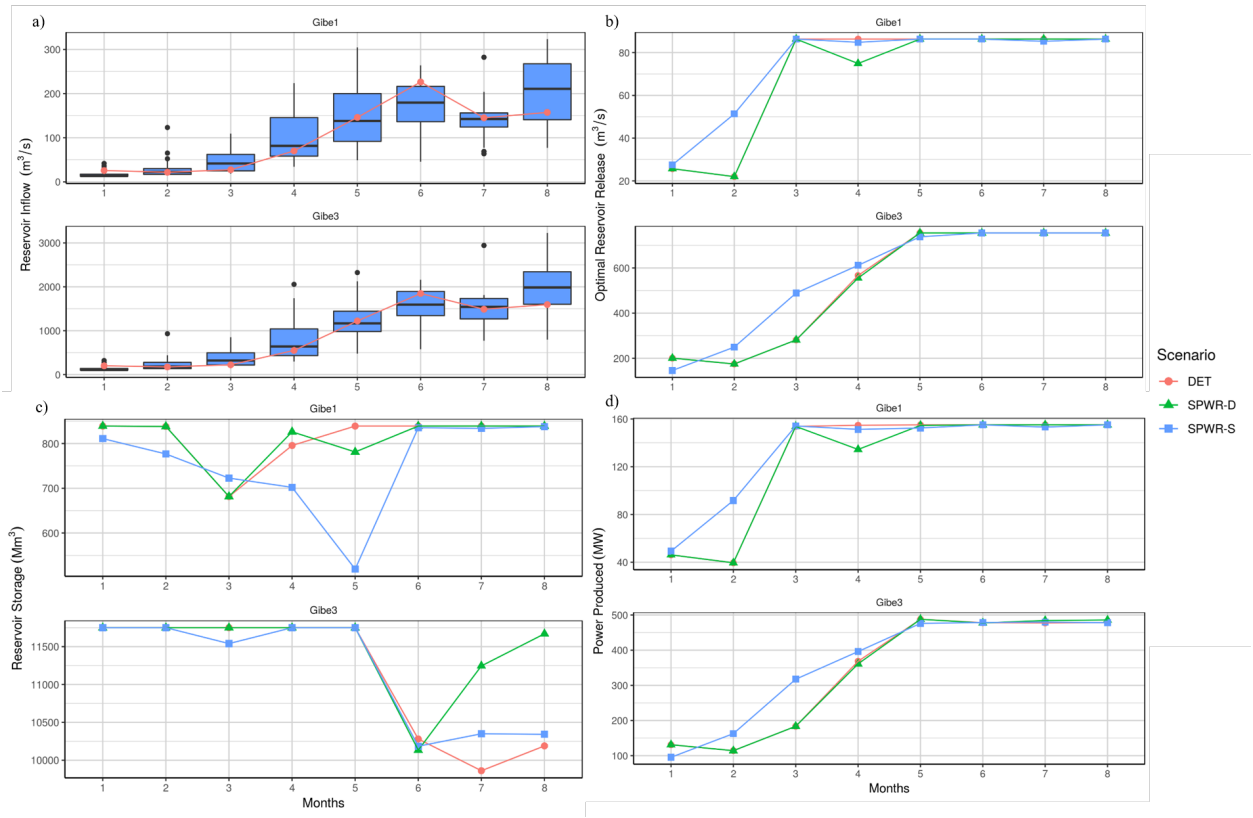

Figure 4.8: a) Reservoir inflows, b) optimal release decisions, c) storage and d) power produced in the Gibe 1 and Gibe 3 reservoirs for the 8 month planning horizon (February 2005 to September 2005) and three scenario structures (DET, SPWR-D and SPWR-S). Note: the results correspond to the first stage of each iteration of the deterministic and stochastic programming with recourse model.

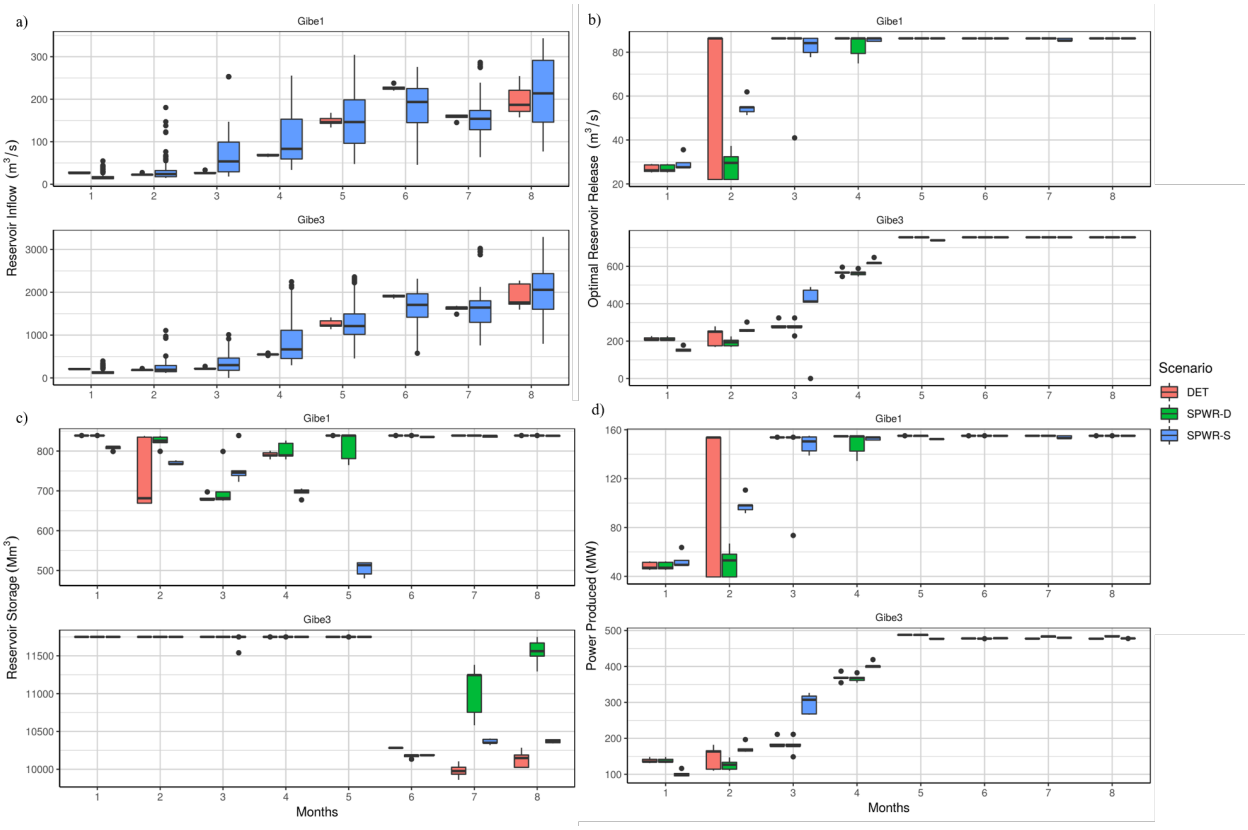

Figure 4.9: Uncertainty in a) reservoir inflows, b) optimal release decisions, c) storage and d) power produced in the Gibe 1 and Gibe 3 reservoirs for the 8 month planning horizon (February 2005 to September 2005) and three scenario structures (DET, SPWR-D and SPWR-S), due to uncertainty in model parameters derived from Bayesian calibration. Note: the results correspond to the first stage of each iteration of the deterministic and stochastic programming with recourse model.

Table 4.2: Different measures of absolute and relative dispersion determined for the ensemble inflows into Gibe I and Gibe III reservoirs generated using the best performing Noah-MP parameter set

Month	**Range (m3/s)**	**Standard Deviation (m3/s)**	**Coefficient of Variation**	**Coefficient of Range**
February	(29.9, 226.1)	(7.7, 58.8)	(0.45, 0.43)	(0.56, 0.54)
March	(108.3, 815.8)	(21.6, 160.9)	(0.74, 0.69)	(0.79, 0.78)
April	(91.0, 681.4)	(25.8, 184.2)	(0.55, 0.49)	(0.71, 0.67)
May	(189.5, 1760.4)	(59.4, 498.1)	(0.59, 0.60)	(0.73, 0.75)
June	(255.5, 1847.1)	(64.5, 456.7)	(0.45, 0.37)	(0.72, 0.66)
July	(218.3, 1583.8)	(54.0, 390.8)	(0.31, 0.24)	(0.71, 0.58)
August	(218.6, 2172.9)	(53.1, 512.1)	(0.37, 0.34)	(0.63, 0.59)
September	(246.5, 2430.8)	(81.4, 570.1)	(0.39, 0.29)	(0.62, 0.60)

a In the parenthesis, values for Gibe I inflows are presented first, followed by values for Gibe III inflows

Table 4.3: Different measures of absolute and relative dispersion determined for the ensemble inflows into Gibe I and Gibe III reservoirs generated using all the behavioral Noah-MP parameter sets

Month	**Range (m3/s)**	**Standard Deviation (m3/s)**	**Coefficient of Variation**	**Coefficient of Range**
February	(43.1, 301.8)	(8.2, 61.9)	(0.46, 0.43)	(0.65, 0.62)
March	(165.7, 990.8)	(24.9, 168.5)	(0.79, 0.68)	(0.85, 0.81)
April	(234.8, 1009.0)	(85.3, 235.8)	(0.94, 0.74)	(0.87, 0.71)
May	(221.9, 1949.9)	(60.9, 508.0)	(0.58, 0.60)	(0.77, 0.77)
June	(256.9, 1904.9)	(62.6, 454.1)	(0.42, 0.35)	(0.73, 0.68)
July	(230.3, 1740.2)	(53.4, 399.6)	(0.29, 0.24)	(0.72, 0.60)
August	(223.0, 2264.7)	(50.2, 510.4)	(0.33, 0.32)	(0.63, 0.60)
September	(266.4, 2494.6)	(80.5, 557.9)	(0.38, 0.28)	(0.63, 0.61)

a In the parenthesis, values for Gibe I inflows are presented first, followed by values for Gibe III inflows

4.5 Conclusions and future work

In this study, we developed a framework for seasonal hydropower planning in regions where reliable streamflow measurements are unavailable. Within this framework, we investigated the potential of combining seasonal precipitation forecasts from NWP models, ET-calibrated spatially distributed hydrologic models, and stochastic programming with recourse models for optimizing hydropower production in data-scarce regions. We compared three different inflow scenario structures that combine deterministic and stochastic reservoir inflow forecasts (DET, SPWR-D, SPWR-S). In addition, we used a Bayesian calibration approach to quantify the impact of hydrologic model parameter uncertainty on the reservoir inflow forecasts, optimal release decisions, and the hydropower production. We applied the framework to a cascade of two reservoirs in the Omo-Gibe river basin using NMME seasonal forecasts, the Noah-MP hydrologic model, and a deterministic and stochastic programming with recourse model. We draw the following conclusions - 1) The NWP-based 3-model, 30-member ensemble precipitation forecasts from NMME are accurate at short lead times (1-3 months), 2) Bayesian model averaging (BMA) of the ensemble precipitation forecasts outperform the ensemble mean as well as the individual models at all lead times, 3) The Noah-MP model calibrated with remote sensing-based ET estimates (GLEAM) performs well in simulating evapotranspiration in the calibration and validation time periods, 4) The ensemble seasonal inflow forecasts exhibit considerable uncertainty, with the deterministic inflow values (DET scenario structure) being consistently lower than the mean of the raw ensembles, 5) The uncertainty in the inflow forecasts affect the optimized release decisions only in the dry months of the study period. This finding agrees with previous studies in which accurate streamflow forecasts improve system performance mainly in dry situations such as droughts (Turner et al., 2017), 6) In the wet period all the ensemble reservoir inflows are very high and the optimized power releases are governed by the capacity of the power plants, 7) In terms of hydropower production, SPWR-D scenario structure leads to the most conservative estimate (994 MW), followed by DET (1,014) and SPWR-S (1,060), 8) Model parameter uncertainty has significant impact (200-250 MW) on the optimized hydropower results, and 9),

using inflow forecasts and stochastic programming, there is a substantial gain in hydropower production.

Currently, the NWP-based seasonal precipitation forecasts are available for 8-12 months. This hinders the applicability of the framework in seasonal hydropower planning which requires forecasts with lead times greater than 12 months. Therefore, future work involves extending the seasonal forecasts beyond lead times of 12 months. In our study, we assume that evapotranspiration is a reliable proxy for streamflow in calibrating and validating the hydrologic model. Although this is a valid assumption based on existing calibration literature, it would be interesting to analyze the impact of using other fluxes (such as soil moisture and total water storage) as proxies for streamflow. The framework developed in this study for seasonal hydropower planning, provides the flexibility of testing different seasonal precipitation forecasts, forecast post-processing methods other BMA, different hydrologic models, and optimization algorithms. Finally, future work also involves the application of the developed methodology to other study regions with larger hydropower reservoir systems.

CHAPTER 5

Dissertation Conclusions and Future Work

The main objective of the dissertation was to improve hydrologic and hydropower studies in data-scarce regions using satellite-based remote sensing. In this regard, the dissertation has led to three original contributions: 1) a validation framework for remotely sensed precipitation and evapotranspiration datasets without ground-based measurements, 2) a framework for calibrating large-scale hydrologic models with multiple flux measurements, and 3) a general methodology for seasonal hydropower planning in data-scarce regions using remote sensing data, multi-model ensemble precipitation forecasts, stochastic programming with recourse models. In this section, a summary of the conclusions drawn from the results presented in chapters two, three, and four are presented. Finally, possible improvements in the presented research are suggested. In addition, potential lines of inquiry which can be pursued to improve hydrologic research in data-scarce regions is discussed.

5.1 Conclusions and original contributions

In Chapter 2, the primary impediment of using satellite-based remote sensing data for hydrologic studies in data-scarce regions i.e., the large uncertainty in the available remotely sensed hydrologic fluxes, was addressed. A novel approach to test the consistency of precipitation and evapotranspiration datasets using a parsimonious combined water-energy balance model (Budyko hypothesis) was proposed. This approach is conceptually different from methods such as triple collocation (Stoffelen, 1998; McColl et al., 2014), where the focus is on statistically understanding the errors in datasets using at least three independent estimates of the target variable. The approach is also different from a probabilistic approach wherein all

available datasets are used to quantify the uncertainty in the final results. The approach proposed in the dissertation was shown to mimic traditional validation methodologies which use ground-based measurements. The application of the framework to a data-scarce region (Omo-Gibe River basin) revealed the importance of ranking and selecting the most accurate precipitation and evapotranspiration datasets before using them in hydrologic studies.

Chapter 3 addressed a fundamental question in hydrologic modeling of data-scarce catchments: How can the wealth of remotely sensed estimates of hydrologic fluxes and storages be meaningfully used to constrain parameters of large scale hydrologic models? The novel approach combined Bayesian and Pareto-optimal approaches to understand the trade-offs among accurate simulation of multiple fluxes and the factors governing such trade-offs. The results reveal several new insights into the value of different hydrologic data such as ET, SM, and SF in improving the realism of hydrologic models. Contrary to previous studies, the results showed that significant trade-offs exist among accurate simulations of different fluxes by hydrologic models. In addition, the framework enables comprehensive diagnosis of model parameters. In contrast to traditional parameter sensitivity, the importance of different model parameters for accurately accurately simulating multiple parameters simultaneously is highlighted. For data-scarce regions, it is seen that evapotranspiration is a better proxy for streamflow compared to soil moisture. This result has significant implications on model calibration in regions where streamflow measurements are not available.

The conclusions drawn in Chapter 2 (regarding the best precipitation and evapotranspiration dataset) and Chapter 3 (regarding the best proxy for streamflow in calibrating hydrologic models) were used to develop a seasonal hydropower planning for data-scarce regions in Chapter 4. Inspite of the lack of skill in seasonal forecasts at long lead times, the research represented the first application of ensemble NWP-based global seasonal precipitation forecasts for hydropower planning. The comparison of different optimization models (deterministic, traditional stochastic programming with recourse, reformulated stochastic programming with recourse) highlighted the importance taking into account input uncertainty in hydropower optimization. In addition, the use of Bayesian calibration enabled

the quantification of the value of incorporating model parameter uncertainty in hydropower planning.

5.2 Future work

Data-scarce regions present significant challenges for conducting robust studies concerning either their hydrology or water resources systems. Although the dissertation addresses a number of important problems, several important issues persist. In addressing these challenges, the role of satellite-based is important. Several different approaches, which have not been investigated in this dissertation, such as data assimilation needs to be explored to improve the hydrologic understanding of data-scarce catchments. However the potential improvements and future work discussed below is restricted to the research carried out in this dissertation:

- The Budyko hypothesis used in developing the validation framework (Chapter 1) is only applicable for long-term time scales. Future work involves extending the framework to account for inter-annual and intra-annual variations in precipitation and evapotranspiration.

- The general validation framework (Chapter 1) can be used for global validation of precipitation and evapotranspiration datasets as it does not need ground-based measurements.

- The calibration framework developed in Chapter 2 was applied for evapotranspiration, soil moisture, and streamflow variables. Future work involved extending the analysis to other variables such snow water equivalent (SWE) and total water storage (TWS).

- The calibration framework can be used for model intercomparison studies to understand the trade-offs in accuracies and understand parameter deficiencies in different large scale hydrologic models and different study regions.

- The seasonal hydropower planning framework developed in Chapter 3 needs to be operationalized in the study region (Omo-Gibe river basin). Additionaly, future work involves extending the hydropower optimization methodology to sub-seasonal and real-time temporal scales.

BIBLIOGRAPHY

Adler, R. F., Kidd, C., Petty, G., Morissey, M., and Goodman, H. M. (2001). Intercomparison of global precipitation products: The third precipitation intercomparison project (pip - 3). *Bulletin of the American Meteorological Society*, 82(7):1377–1396.

Alemu, E. T., Palmer, R. N., Polebitski, A., and Meaker, B. (2011). Decision support system for optimizing reservoir operations using ensemble streamflow predictions. *Journal of Water Resources Planning and Management*, 137(1):72–82.

Alvarez-Garreton, C., Ryu, D., Western, A. W., Crow, W. T., Su, C.-H., and Robertson, D. R. (2016). Dual assimilation of satellite soil moisture to improve streamflow prediction in data-scarce catchments. *Water Resources Research*, 52(7):5357–5375.

Ashouri, H., Hsu, K.-L., Sorooshian, S., Braithwaite, D. K., Knapp, K. R., Cecil, L. D., Nelson, B. R., and Prat, O. P. (2015). Persiann-cdr: Daily precipitation climate data record from multisatellite observations for hydrological and climate studies. *Bulletin of the American Meteorological Society*, 96(1):69–83.

Ball, J. T., Woodrow, I. E., and Berry, J. A. (1987). *A Model Predicting Stomatal Conductance and its Contribution to the Control of Photosynthesis under Different Environmental Conditions*, pages 221–224. Springer Netherlands, Dordrecht.

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A. I. J. M., Weedon, G. P., Brocca, L., Pappenberger, F., Huffman, G. J., and Wood, E. F. (2017). Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrology and Earth System Sciences*, 21(12):6201–6217.

Becker, E., den Dool, H. v., and Zhang, Q. (2014). Predictability and forecast skill in nmme. *Journal of Climate*, 27(15):5891–5906.

Becker, E. and van den Dool, H. (2016). Probabilistic seasonal forecasts in the north american multimodel ensemble: A baseline skill assessment. *Journal of Climate*, 29(8):3015–3026.

Beven, K. (1996). Equifinality and uncertainty in geomorphological modelling. In *The Scientific Nature of Geomorphology: Proceedings of the 27th Binghamton Symposium in Geomorphology, Held 27-29 September, 1996*, volume 27, page 289. John Wiley & Sons.

Beven, K. (2001). How far can we go in distributed hydrological modelling? *Hydrology and Earth System Sciences*, 5(1):1–12.

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, 320(1):18 – 36. The model parameter estimation experiment.

Beven, K. and Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, 6(3):279–298.

Beven, K. and Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Journal*, 24(1):43–69.

Beven, K. J. (2018). On hypothesis testing in hydrology: Why falsification of models is still a really good idea. *Wiley Interdisciplinary Reviews: Water*, 5(3):e1278.

Bitew, M. and Gebremichael, M. (2011). Assessment of satellite rainfall products for streamflow simulation in medium watersheds of the ethiopian highlands. *Hydrology and Earth System Sciences*, 15(4):1147–1155.

Block, P. (2011). Tailoring seasonal climate forecasts for hydropower operations. *Hydrology and Earth System Sciences*, 15(4):1355–1368.

Block, P. J., Souza Filho, F. A., Sun, L., and Kwon, H.-H. (2009). A streamflow forecasting framework using multiple climate and hydrological models1. *JAWRA Journal of the American Water Resources Association*, 45(4):828–843.

Borovikov, A., Cullather, R., Kovach, R., Marshak, J., Vernieres, G., Vikhliaev, Y., Zhao, B., and Li, Z. (2017). Geos-5 seasonal forecast system. *Climate Dynamics*.

Budyko, M. (1974). *Climate and Life*. International geophysics series. Academic Press.

Cai, X., Yang, Z.-L., David, C. H., Niu, G.-Y., and Rodell, M. (2014). Hydrological evaluation of the noah-mp land surface model for the mississippi river basin. *Journal of Geophysical Research: Atmospheres*, 119(1):23–38. 2013JD020792.

Carmona, A., Poveda, G., Sivapalan, M., Vallejo-Bernal, S., and Bustamante, E. (2016). A scaling approach to budyko's framework and the complementary relationship of evapotranspiration in humid environments: case study of the amazon river basin. *Hydrology and Earth System Sciences*, 20(2):589–603.

Cash, B. A., Manganello, J. V., and Kinter, J. L. (2017). Evaluation of nmme temperature and precipitation bias and forecast skill for south asia. *Climate Dynamics*.

Chaemiso, S. E., Abebe, A., and Pingale, S. M. (2016). Assessment of the impact of climate change on surface hydrological processes using swat: a case study of omo-gibe river basin, ethiopia. *Modeling Earth Systems and Environment*, 2(4):1–15.

Chang, X., Liu, X., and Zhou, W. (2010). Hydropower in china at present and its further development. *Energy*, 35(11):4400 – 4406. Energy and Its Sustainable Development for China.

Chen, F., Janjić, Z., and Mitchell, K. (1997). Impact of atmospheric surface-layer parameterizations in the new land-surface scheme of the ncep mesoscale eta model. *Boundary-Layer Meteorology*, 85(3):391–421.

Chen, S., Liu, H., You, Y., Mullens, E., Hu, J., Yuan, Y., Huang, M., He, L., Luo, Y., Zeng, X., Tang, G., and Hong, Y. (2014). Evaluation of high-resolution precipitation estimates from satellites during july 2012 beijing flood event using dense rain gauge observations. *PLoS ONE*, 9(4):1–7.

Choudhury, B. (1999). Evaluation of an empirical equation for annual evaporation using field observations and results from a biophysical model. *Journal of Hydrology*, 216(12):99 – 110.

Christensen, N. S. and Lettenmaier, D. P. (2007). A multimodel ensemble approach to assessment of climate change impacts on the hydrology and water resources of the colorado river basin. *Hydrology and Earth System Sciences*, 11(4):1417–1434.

Conway, D., Dalin, C., Landman, W. A., and Osborn, T. J. (2017). Hydropower plans in eastern and southern africa increase risk of concurrent climate-related electricity supply disruption. *Nature Energy*, 2(12):946–953.

Conway, D., van Garderen, E. A., Deryng, D., Dorling, S., Krueger, T., Landman, W., Lankford, B., Lebek, K., Osborn, T., Ringler, C., Thurlow, J., Zhu, T., and Dalin, C. (2015). Climate and southern africa's water-energy-food nexus. *Nature Climate Change*, 5:837 EP –. Review Article.

Cosby, B. J., Hornberger, G. M., Clapp, R. B., and Ginn, T. R. (1984). A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils. *Water Resources Research*, 20(6):682–690.

Cuntz, M., Mai, J., Samaniego, L., Clark, M., Wulfmeyer, V., Branch, O., Attinger, S., and Thober, S. (2016). The impact of standard and hard-coded parameters on the hydrologic fluxes in the noah-mp land surface model. *Journal of Geophysical Research: Atmospheres*, 121(18):10,676–10,700. 2016JD025097.

Cuo, L., Zhang, Y., Gao, Y., Hao, Z., and Cairang, L. (2013). The impacts of climate change and land cover/use transition on the hydrology in the upper yellow river basin, china. *Journal of Hydrology*, 502(Supplement C):37 – 52.

Day, G. N. (1985). Extended streamflow forecasting using nwsrfs. *Journal of Water Resources Planning and Management*, 111(2):157–170.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197.

Demirel, M. C., Booij, M. J., and Hoekstra, A. Y. (2013). Effect of different uncertainty sources on the skill of 10 day ensemble low flow forecasts for two hydrological models. *Water Resources Research*, 49(7):4035–4053.

Derber, J. C., Parrish, D. F., and Lord, S. J. (1991). The new global operational analysis system at the national meteorological center. *Weather and Forecasting*, 6(4):538–547.

Donohue, R. J., Roderick, M. L., and McVicar, T. R. (2012). Roots, storms and soil pores: Incorporating key ecohydrological processes into budyko hydrological model. *Journal of Hydrology*, 436437:35 – 50.

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A., Haas, E., Hamer, P. D., Hirschi, M., Ikonen, J., de Jeu, R., Kidd, R., Lahoz, W., Liu, Y. Y., Miralles, D., Mistelbauer, T., Nicolai-Shaw, N., Parinussa, R., Pratola, C., Reimer, C., van der Schalie, R., Seneviratne, S. I., Smolander, T., and Lecomte, P. (2017). Esa cci soil moisture for improved earth system understanding: State-of-the art and future directions. *Remote Sensing of Environment*, 203(Supplement C):185 – 215.

Eden, J. M., van Oldenborgh, G. J., Hawkins, E., and Suckling, E. B. (2015). A global empirical system for probabilistic seasonal climate prediction. *Geoscientific Model Development*, 8(12):3947–3973.

Efstratiadis, A. and Koutsoyiannis, D. (2010). One decade of multi-objective calibration approaches in hydrological modelling: a review. *Hydrological Sciences Journal*, 55(1):58–78.

Etkin, D., Kirshen, P., Watkins, D., Roncoli, C., Sanon, M., Some, L., Dembele, Y., Sanfo, J., Zoungrana, J., and Hoogenboom, G. (2015). Stochastic programming for improved multiuse reservoir operation in burkina faso, west africa. *Journal of Water Resources Planning and Management*, 141(3):04014056.

Fenicia, F., Savenije, H. H. G., Matgen, P., and Pfister, L. (2007). A comparison of alternative multiobjective calibration strategies for hydrological modeling. *Water Resources Research*, 43(3).

Finnis, J., Hsieh, W. W., Lin, H., and Merryfield, W. J. (2012). Non-linear post-processing of numerical seasonal climate forecasts. *Atmosphere-Ocean*, 50(2):207–218.

Fu, B. (1981). On the calculation of the evaporation from land surface. *Sci. Atmos. Sin*, 5(1):23–31.

Gado Djibo, A., Karambiri, H., Seidou, O., Sittichok, K., Paturel, J. E., and Saley, H. M. (2015). Statistical seasonal streamflow forecasting using probabilistic approach over west african sahel. *Natural Hazards*, 79(2):699–722.

GAMS (2018). General algebraic modeling system. http://www.gams.com.

Gebregiorgis, A. and Hossain, F. (2014). Making satellite precipitation data work for the developing world. *IEEE Geoscience and Remote Sensing Magazine*, 2(2):24–36.

Gentine, P., D'Odorico, P., Lintner, B. R., Sivandran, G., and Salvucci, G. (2012). Interdependence of climate, soil, and vegetation as constrained by the budyko curve. *Geophysical Research Letters*, 39(19):n/a–n/a. L19404.

Gerlitz, L., Vorogushyn, S., Apel, H., Gafurov, A., Unger-Shayesteh, K., and Merz, B. (2016). A statistically based seasonal precipitation forecast model with automatic predictor selection and its application to central and south asia. *Hydrology and Earth System Sciences*, 20(11):4605–4623.

Greve, P., Gudmundsson, L., Orlowsky, B., and Seneviratne, S. I. (2015). Introducing a probabilistic budyko framework. *Geophysical Research Letters*, 42(7):2261–2269.

Greve, P., Gudmundsson, L., Orlowsky, B., and Seneviratne, S. I. (2016). A two-parameter budyko function to represent conditions under which evapotranspiration exceeds precipitation. *Hydrology and Earth System Sciences*, 20(6):2195.

Greve, P., Orlowsky, B., Mueller, B., Sheffield, J., Reichstein, M., and Seneviratne, S. I. (2014). Global assessment of trends in wetting and drying over land. *Nature Geoscience*, 7(10):716–721.

Gupta, H. V., Bastidas, L. A., Sorooshian, S., Shuttleworth, W. J., and Yang, Z. L. (1999). Parameter estimation of a land surface scheme using multicriteria methods. *Journal of Geophysical Research: Atmospheres*.

Gupta, H. V., Sorooshian, S., and Yapo, P. O. (1998). Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research*, 34(4):751–763.

Gupta, H. V., Wagener, T., and Liu, Y. (2008). Reconciling theory with observations: elements of a diagnostic approach to model evaluation. *Hydrological Processes*, 22(18):3802–3813.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242.

Herr, H. D. and Krzysztofowicz, R. (2015). Ensemble bayesian forecasting system part i: Theory and algorithms. *Journal of Hydrology*, 524:789 – 802.

Hirpa, F. A., Gebremichael, M., and Hopson, T. (2010). Evaluation of high-resolution satellite precipitation products over very complex terrain in ethiopia. *Journal of Applied Meteorology and Climatology*, 49(5):1044–1051.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and e. i. george, and a rejoinder by the authors. *Statist. Sci.*, 14(4):382–417.

Hogue, T. S., Bastidas, L. A., Gupta, H. V., and Sorooshian, S. (2006). Evaluating model performance and parameter behavior for varying levels of land surface model complexity. *Water Resources Research*, 42(8).

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., Hong, Y., Bowman, K. P., and Stocker, E. F. (2007). The trmm multisatellite precipitation analysis (tmpa): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of Hydrometeorology*, 8(1):38–55.

Immerzeel, W. and Droogers, P. (2008). Calibration of a distributed hydrological model based on satellite evapotranspiration. *Journal of Hydrology*, 349(3):411 – 424.

Istanbulluoglu, E., Wang, T., Wright, O. M., and Lenters, J. D. (2012). Interpretation of hydrologic trends from a water balance perspective: The role of groundwater storage in the budyko hypothesis. *Water Resources Research*, 48(3):n/a–n/a. W00H16.

Jia, X., Lin, H., and Derome, J. (2010). Improving seasonal forecast skill of north american surface air temperature in fall using a postprocessing method. *Monthly Weather Review*, 138(5):1843–1857.

Jordan, R. (1991). A one-dimensional temperature model for a snow cover: Technical documentation for sntherm. 89. Technical report, Cold Regions Research and Engineering Lab Hanover NH.

Joyce, R. J., Janowiak, J. E., Arkin, P. A., and Xie, P. (2004). Cmorph: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *Journal of Hydrometeorology*, 5(3):487–503.

Kavetski, D., Kuczera, G., and Franks, S. W. (2006). Bayesian analysis of input uncertainty in hydrological modeling: 1. theory. *Water Resources Research*, 42(3).

Kennedy, J. and Eberhart, R. C. (2001). *Swarm Intelligence*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Khanmohammadi, N., Rezaie, H., Montaseri, M., and Behmanesh, J. (2018). Regional probability distribution of the annual reference evapotranspiration and its effective parameters in iran. *Theoretical and Applied Climatology*, 134(1):411–422.

Khu, S. T. and Madsen, H. (2005). Multiobjective calibration with pareto preference ordering: An application to rainfall-runoff model calibration. *Water Resources Research*, 41(3).

Kidd, C. and Huffman, G. (2011). Global precipitation measurement. *Meteorological Applications*, 18(3):334–353.

Kim, Y.-O. and Palmer, R. N. (1997). Value of seasonal flow forecasts in bayesian stochastic programming. *Journal of Water Resources Planning and Management*, 123(6):327–335.

Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., Paolino, D. A., Zhang, Q., van den Dool, H., Saha, S., Mendez, M. P., Becker, E., Peng, P., Tripp, P., Huang, J., DeWitt, D. G., Tippett, M. K., Barnston, A. G., Li, S., Rosati, A., Schubert, S. D., Rienecker, M., Suarez, M., Li, Z. E., Marshak, J., Lim, Y.-K., Tribbia, J., Pegion, K., Merryfield, W. J., Denis, B., and Wood, E. F. (2014). The north american multimodel ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bulletin of the American Meteorological Society*, 95(4):585–601.

Koppa, A. and Gebremichael, M. (2017). A framework for validation of remotely sensed precipitation and evapotranspiration based on the budyko hypothesis. *Water Resources Research*, 53(10):8487–8499.

Koster, R. D., Mahanama, S. P. P., Livneh, B., Lettenmaier, D. P., and Reichle, R. H. (2010). Skill in streamflow forecasts derived from large-scale estimates of soil moisture and snow. *Nature Geoscience*, 3:613 EP –.

Krishnamurti, T. N., Kumar, V., Simon, A., Bhardwaj, A., Ghosh, T., and Ross, R. (2016). A review of multimodel superensemble forecasting for weather, seasonal climate, and hurricanes. *Reviews of Geophysics*, 54(2):336–377.

Krishnaswamy, J., Vaidyanathan, S., Rajagopalan, B., Bonell, M., Sankaran, M., Bhalla, R. S., and Badiger, S. (2015). Non-stationary and non-linear influence of enso and indian

ocean dipole on the variability of indian monsoon rainfall and extreme rain events. *Climate Dynamics*, 45(1):175–184.

Kumar, A., Chen, M., and Wang, W. (2013). Understanding prediction skill of seasonal mean precipitation over the tropics. *Journal of Climate*, 26(15):5674–5681.

Kumar, S., Peters-Lidard, C., Tian, Y., Houser, P., Geiger, J., Olden, S., Lighty, L., Eastman, J., Doty, B., Dirmeyer, P., Adams, J., Mitchell, K., Wood, E., and Sheffield, J. (2006). Land information system: An interoperable framework for high resolution land surface modeling. *Environmental Modelling & Software*, 21(10):1402 – 1415.

Labadie, J. W. (2004). Optimal operation of multireservoir systems: State-of-the-art review. *Journal of Water Resources Planning and Management*, 130(2):93–111.

Lakshmi, V. (2004). The role of satellite remote sensing in the prediction of ungauged basins. *Hydrological Processes*, 18(5):1029–1034.

Laloy, E., Rogiers, B., Vrugt, J. A., Mallants, D., and Jacques, D. (2013). Efficient posterior exploration of a high-dimensional groundwater model from two-stage markov chain monte carlo simulation and polynomial chaos expansion. *Water Resources Research*, 49(5):2664–2682.

Lavers, D., Luo, L., and Wood, E. F. (2009). A multiple model assessment of seasonal climate forecast skill for applications. *Geophysical Research Letters*, 36(23).

Lee, Y., Kim, S.-K., and Ko, I. H. (2008). Multistage stochastic linear programming model for daily coordinated multi-reservoir operation. *Journal of Hydroinformatics*, 10(1):23–41.

Lehner, B., Verdin, K., and Jarvis, A. (2006). Hydrosheds technical documentation. *World Wildlife Fund US, Washington DC. Available at http://hydrosheds.cr.usgs.gov*.

Leng, G., Tang, Q., and Rayburg, S. (2015). Climate change impacts on meteorological, agricultural and hydrological droughts in china. *Global and Planetary Change*, 126(Supplement C):23 – 34.

Lettenmaier, D. P., Alsdorf, D., Dozier, J., Huffman, G. J., Pan, M., and Wood, E. F. (2015). Inroads of remote sensing into hydrologic science during the wrr era. *Water Resources Research*, 51(9):7309–7342.

Li, D., Pan, M., Cong, Z., Zhang, L., and Wood, E. (2013). Vegetation control on water and energy balance within the budyko framework. *Water Resources Research*, 49(2):969–976.

Li, D., Wrzesien, M. L., Durand, M., Adam, J., and Lettenmaier, D. P. (2017). How much runoff originates as snow in the western united states, and how will that change in the future? *Geophysical Research Letters*, 44(12):6163–6172. 2017GL073551.

López López, P., Sutanudjaja, E. H., Schellekens, J., Sterk, G., and Bierkens, M. F. P. (2017). Calibration of a large-scale hydrological model using satellite-based soil moisture and evapotranspiration products. *Hydrology and Earth System Sciences*, 21(6):3125–3144.

Lorenz, C. and Kunstmann, H. (2012). The hydrological cycle in three state-of-the-art reanalyses: Intercomparison and performance analysis. *Journal of Hydrometeorology*, 13(5):1397–1420.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2):130–141.

Lund, J. (1996). Developing seasonal and long-term reservoir system operation plans using hec-prm. Technical report, Hydrologic Engineering Center Davis CA.

Ma, N., Niu, G.-Y., Xia, Y., Cai, X., Zhang, Y., Ma, Y., and Fang, Y. (2017). A systematic evaluation of noah-mp in simulating land-atmosphere energy, water, and carbon exchanges over the continental united states. *Journal of Geophysical Research: Atmospheres*, 122(22):12,245–12,268.

Madsen, H. (2003). Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. *Advances in Water Resources*, 26(2):205 – 216.

Mahanama, S. P., Koster, R. D., Reichle, R. H., and Zubair, L. (2008). The role of soil moisture initialization in subseasonal and seasonal streamflow prediction a case study in sri lanka. *Advances in Water Resources*, 31(10):1333 – 1343.

Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Férnandez-Prieto, D., Beck, H. E., Dorigo, W. A., and Verhoest, N. E. C. (2016). Gleam v3: satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development Discussions*, 2016:1–36.

McColl, K. A., Vogelzang, J., Konings, A. G., Entekhabi, D., Piles, M., and Stoffelen, A. (2014). Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target. *Geophysical Research Letters*, 41(17):6229–6236.

Mendoza, P. A., Clark, M. P., Mizukami, N., Newman, A. J., Barlage, M., Gutmann, E. D., Rasmussen, R. M., Rajagopalan, B., Brekke, L. D., and Arnold, J. R. (2015). Effects of hydrologic model choice and calibration on the portrayal of climate change impacts. *Journal of Hydrometeorology*, 16(2):762–780.

Merryfield, W. J., Lee, W.-S., Boer, G. J., Kharin, V. V., Scinocca, J. F., Flato, G. M., Ajayamohan, R. S., Fyfe, J. C., Tang, Y., and Polavarapu, S. (2013). The canadian seasonal to interannual prediction system. part i: Models and initialization. *Monthly Weather Review*, 141(8):2910–2945.

Middelkoop, H., Daamen, K., Gellens, D., Grabs, W., Kwadijk, J. C. J., Lang, H., Parmet, B. W. A. H., Schädler, B., Schulla, J., and Wilke, K. (2001). Impact of climate change on hydrological regimes and water resources management in the rhine basin. *Climatic Change*, 49(1):105–128.

Miralles, D., Holmes, T., De Jeu, R., Gash, J., Meesters, A., and Dolman, A. (2011). Global land-surface evaporation estimated from satellite-based observations. *Hydrology and Earth System Sciences*, 15(2):453.

Miralles, D. G., Jiménez, C., Jung, M., Michel, D., Ershadi, A., McCabe, M. F., Hirschi, M., Martens, B., Dolman, A. J., Fisher, J. B., Mu, Q., Seneviratne, S. I., Wood, E. F., and Fernández-Prieto, D. (2016). The wacmos-et project – part-2: Evaluation of global terrestrial evaporation data sets. *Hydrology and Earth System Sciences*, 20(2):823–842.

Moran, E. F., Lopez, M. C., Moore, N., Müller, N., and Hyndman, D. W. (2018). Sustainable hydropower in the 21st century. *Proceedings of the National Academy of Sciences*, 115(47):11891–11898.

Mu, Q., Heinsch, F. A., Zhao, M., and Running, S. W. (2007). Development of a global evapotranspiration algorithm based on {MODIS} and global meteorology data. *Remote Sensing of Environment*, 111(4):519 – 536.

Niu, G.-Y. and Yang, Z.-L. (2006). Effects of frozen soil on snowmelt runoff and soil water storage at a continental scale. *Journal of Hydrometeorology*, 7(5):937–952.

Niu, G.-Y., Yang, Z.-L., Dickinson, R. E., Gulden, L. E., and Su, H. (2007). Development of a simple groundwater model for use in climate models and evaluation with gravity recovery and climate experiment data. *Journal of Geophysical Research: Atmospheres*, 112(D7).

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., Tewari, M., and Xia, Y. (2011). The community noah land surface model with multiparameterization options (noah-mp): 1. model description and evaluation with local-scale measurements. *Journal of Geophysical Research: Atmospheres*, 116(D12):n/a–n/a. D12109.

O'Connor, J. E., Duda, J. J., and Grant, G. E. (2015). 1000 dams down and counting. *Science*, 348(6234):496–497.

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Dlcluse, P., Dqu, M., Dez, E., Doblas-Reyes, F. J., Feddersen, H., Graham, R., Gualdi, S., Gurmy, J.-F., Hagedorn, R., Hoshen, M., Keenlyside, N., Latif, M., Lazar, A., Maisonnave, E., Marletto,

V., Morse, A. P., Orfila, B., Rogel, P., Terres, J.-M., and Thomson, M. C. (2004). Development of a european multimodel ensemble system for seasonal-to-interannual prediction (demeter). *Bulletin of the American Meteorological Society*, 85(6):853–872.

Piechota, T. C., Chiew, F. H. S., Dracup, J. A., and McMahon, T. A. (1998). Seasonal streamflow forecasting in eastern australia and the el niosouthern oscillation. *Water Resources Research*, 34(11):3035–3044.

Porporato, A., Daly, E., and Rodriguez-Iturbe, I. (2004). Soil water balance and ecosystem response to climate change. *The American Naturalist*, 164(5):625–632.

Quiring, S. M., Ford, T. W., Wang, J. K., Khong, A., Harris, E., Lindgren, T., Goldberg, D. W., and Li, Z. (2016). The north american soil moisture database: Development and applications. *Bulletin of the American Meteorological Society*, 97(8):1441–1459.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174.

Rakovec, O., Kumar, R., Attinger, S., and Samaniego, L. (2016a). Improving the realism of hydrologic model functioning through multivariate parameter estimation. *Water Resources Research*, 52(10):7779–7792.

Rakovec, O., Kumar, R., Mai, J., Cuntz, M., Thober, S., Zink, M., Attinger, S., Schfer, D., Schrn, M., and Samaniego, L. (2016b). Multiscale and multivariate evaluation of water fluxes and states over european river basins. *Journal of Hydrometeorology*, 17(1):287–307.

Ratnam, J. V., Behera, S. K., Masumoto, Y., and Yamagata, T. (2014). Remote effects of el nio and modoki events on the austral summer precipitation of southern africa. *Journal of Climate*, 27(10):3802–3815.

Rientjes, T., Muthuwatta, L., Bos, M., Booij, M., and Bhatti, H. (2013). Multi-variable calibration of a semi-distributed hydrological model using streamflow data and satellite-based evapotranspiration. *Journal of Hydrology*, 505:276 – 290.

Romilly, T. G. and Gebremichael, M. (2011). Evaluation of satellite rainfall estimates over ethiopian river basins. *Hydrology and Earth System Sciences*, 15(5):1505. Copyright - Copyright Copernicus GmbH 2011; Last updated - 2012-06-11.

Saavedra Valeriano, O. C., Koike, T., Yang, K., Graf, T., Li, X., Wang, L., and Han, X. (2010). Decision support for dam release during floods using a distributed biosphere hydrological model driven by quantitative precipitation forecasts. *Water Resources Research*, 46(10).

Sadegh, M. and Vrugt, J. A. (2014). Approximate bayesian computation using markov chain monte carlo simulation: Dream(abc). *Water Resources Research*, 50(8):6767–6787.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., Chuang, H.-y., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M. P., van den Dool, H., Zhang, Q., Wang, W., Chen, M., and Becker, E. (2014). The ncep climate forecast system version 2. *Journal of Climate*, 27(6):2185–2208.

Saha, S., Nadiga, S., Thiaw, C., Wang, J., Wang, W., Zhang, Q., Van den Dool, H. M., Pan, H.-L., Moorthi, S., Behringer, D., Stokes, D., Pea, M., Lord, S., White, G., Ebisuzaki, W., Peng, P., and Xie, P. (2006). The ncep climate forecast system. *Journal of Climate*, 19(15):3483–3517.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global sensitivity analysis: the primer*. John Wiley & Sons.

Samaniego, L., Kumar, R., and Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, 46(5).

Samaniego, L., Kumar, R., Thober, S., Rakovec, O., Zink, M., Wanders, N., Eisner, S., Müller Schmied, H., Sutanudjaja, E. H., Warrach-Sagi, K., and Attinger, S. (2017). Toward seamless hydrologic predictions across spatial scales. *Hydrology and Earth System Sciences*, 21(9):4323–4346.

Sankarasubramanian, A. and Vogel, R. M. (2003). Hydroclimatology of the continental united states. *Geophysical Research Letters*, 30(7):n/a–n/a. 1363.

Schepen, A., Wang, Q. J., and Everingham, Y. (2016). Calibration, bridging, and merging to improve gcm seasonal temperature forecasts in australia. *Monthly Weather Review*, 144(6):2421–2441.

Schoups, G. and Vrugt, J. A. (2010). A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-gaussian errors. *Water Resources Research*, 46(10).

Shafii, M., Tolson, B., and Matott, L. S. (2014). Uncertainty-based multi-criteria calibration of rainfall-runoff models: a comparative study. *Stochastic Environmental Research and Risk Assessment*, 28(6):1493–1510.

Sharma, N. K., Tiwari, P. K., and Sood, Y. R. (2013). A comprehensive analysis of strategies, policies and development of hydropower in india: Special emphasis on small hydro power. *Renewable and Sustainable Energy Reviews*, 18:460 – 470.

Sheffield, J., Ferguson, C. R., Troy, T. J., Wood, E. F., and McCabe, M. F. (2009). Closing the terrestrial water budget from satellite remote sensing. *Geophysical Research Letters*, 36(7):n/a–n/a. L07403.

Sheffield, J., Goteti, G., Wen, F., and Wood, E. F. (2004). A simulated soil moisture based drought analysis for the united states. *Journal of Geophysical Research: Atmospheres*, 109(D24):n/a–n/a. D24108.

Shi, X., Wood, A. W., and Lettenmaier, D. P. (2008). How essential is hydrologic model calibration to seasonal streamflow forecasting? *Journal of Hydrometeorology*, 9(6):1350–1363.

Shrestha, D. L., Robertson, D. E., Bennett, J. C., and Wang, Q. J. (2015). Improving precipitation forecasts by generating ensembles through postprocessing. *Monthly Weather Review*, 143(9):3642–3663.

Shukla, S. and Lettenmaier, D. P. (2011). Seasonal hydrologic prediction in the united states: understanding the role of initial hydrologic conditions and seasonal climate forecast skill. *Hydrology and Earth System Sciences*, 15(11):3529–3538.

Siegmund, J., Bliefernicht, J., Laux, P., and Kunstmann, H. (2015). Toward a seasonal precipitation prediction system for west africa: Performance of cfsv2 and high-resolution dynamical downscaling. *Journal of Geophysical Research: Atmospheres*, 120(15):7316–7339.

Sloughter, J. M. L., Raftery, A. E., Gneiting, T., and Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using bayesian model averaging. *Monthly Weather Review*, 135(9):3209–3220.

Snedecor, G. and Cochran, W. (1989). *Statistical Methods*. Iowa University Press.

Souza Filho, F. A. and Lall, U. (2003). Seasonal to interannual ensemble streamflow forecasts for ceara, brazil: Applications of a multivariate, semiparametric algorithm. *Water Resources Research*, 39(11).

Stisen, S. and Sandholt, I. (2010). Evaluation of remote-sensing-based rainfall products through predictive capability in hydrological runoff modelling. *Hydrological Processes*, 24(7):879–891.

Stoffelen, A. (1998). Toward the true near-surface wind speed: Error modeling and calibration using triple collocation. *Journal of Geophysical Research: Oceans*, 103(C4):7755–7766.

Stokstad, E. (1999). Scarcity of rain, stream gages threatens forecasts. *Science*, 285(5431):1199–1200.

Stone, R. (2011). Mayhem on the mekong. *Science*, 333(6044):814–818.

Storn, R. and Price, K. (1997). Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359.

Sutanudjaja, E. H., van Beek, L. P. H., de Jong, S. M., van Geer, F. C., and Bierkens, M. F. P. (2013). Calibrating a large-extent high-resolution coupled groundwater-land surface model using soil moisture and discharge data. *Water Resources Research*, 50(1):687–705.

Sguin, S., Audet, C., and Ct, P. (2017). Scenario-tree modeling for stochastic short-term hydropower operations planning. *Journal of Water Resources Planning and Management*, 143(12):04017073.

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, 106(D7):7183–7192.

Trabucco, A. and Zomer, R. (2009). Global aridity index (global-aridity) and global potential evapo-transpiration (global-pet) geospatial database. *CGIAR Consortium for Spatial Information. Available at http://csi.cgiar.org*.

Trezos, T. and Yeh, W. W.-G. (1987). Use of stochastic dynamic programming for reservoir management. *Water Resources Research*, 23(6):983–996.

Turner, S. W. D., Bennett, J. C., Robertson, D. E., and Galelli, S. (2017). Complex relationship between seasonal streamflow forecast skill and value in reservoir operations. *Hydrology and Earth System Sciences*, 21(9):4841–4859.

Uysal, G., Alvarado-Montero, R., Schwanenberg, D., and ensoy, A. (2018). Real-time flood control by tree-based model predictive control including forecast uncertainty: A case study reservoir in turkey. *Water*, 10(3).

Verseghy, D. L., McFarlane, N. A., and Lazare, M. (1991). Class - a canadian land surface scheme for gcms, ii. vegetation model and coupled runs. *International Journal of Climatology*, 13(4):347–370.

Vogel, R. M. and Wilson, I. (1996). Probability distribution of annual maximum, mean, and minimum streamflows in the united states. *Journal of Hydrologic Engineering*, 1(2):69–76.

Vrugt, J. A. (2016). Markov chain monte carlo simulation using the dream software package: Theory, concepts, and matlab implementation. *Environmental Modelling & Software*, 75:273 – 316.

Vrugt, J. A. and Robinson, B. A. (2007). Improved evolutionary optimization from genetically adaptive multimethod search. *Proc Natl Acad Sci U S A*, 104(3):708–711.

Vrugt, J. A. and Sadegh, M. (2013). Toward diagnostic model calibration and evaluation: Approximate bayesian computation. *Water Resources Research*, 49(7):4335–4345.

Vrugt, J. A., Ter Braak, C., Diks, C., Robinson, B. A., Hyman, J. M., and Higdon, D. (2009a). Accelerating markov chain monte carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Sciences and Numerical Simulation*, 10(3):273–290.

Vrugt, J. A., ter Braak, C. J. F., Clark, M. P., Hyman, J. M., and Robinson, B. A. (2008). Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with markov chain monte carlo simulation. *Water Resources Research*, 44(12).

Vrugt, J. A., ter Braak, C. J. F., Gupta, H. V., and Robinson, B. A. (2009b). Equifinality of formal (dream) and informal (glue) bayesian approaches in hydrologic modeling? *Stochastic Environmental Research and Risk Assessment*, 23(7):1011–1026.

Wanders, N., Bierkens, M. F. P., de Jong, S. M., de Roo, A., and Karssenberg, D. (2014). The benefits of using remotely sensed soil moisture in parameter identification of large-scale hydrological models. *Water Resources Research*, 50(8):6874–6891.

Wang, F., Wang, L., Zhou, H., Saavedra Valeriano, O. C., Koike, T., and Li, W. (2012). Ensemble hydrological prediction-based real-time optimization of a multiobjective reservoir during flood season in a semiarid basin with global numerical weather predictions. *Water Resources Research*, 48(7).

Weisheimer, A., Doblas-Reyes, F. J., Palmer, T. N., Alessandri, A., Arribas, A., Dqu, M., Keenlyside, N., MacVean, M., Navarra, A., and Rogel, P. (2009). Ensembles: A new

multi-model ensemble for seasonal-to-annual predictionsskill and progress beyond demeter in forecasting tropical pacific ssts. *Geophysical Research Letters*, 36(21).

Xu, W., Zhang, C., Peng, Y., Fu, G., and Zhou, H. (2014). A two stage bayesian stochastic optimization model for cascaded hydropower systems considering varying uncertainty of flow forecasts. *Water Resources Research*, 50(12):9267–9286.

Xu, X., Liu, W., Scanlon, B. R., Zhang, L., and Pan, M. (2013). Local and global factors controlling water-energy balances within the budyko framework. *Geophysical Research Letters*, 40(23):6123–6129.

Yang, H., Yang, D., Lei, Z., and Sun, F. (2008). New analytical derivation of the mean annual water-energy balance equation. *Water Resources Research*, 44(3):n/a–n/a. W03410.

Yang, R. and Friedl, M. A. (2003). Modeling the effects of three-dimensional vegetation structure on surface radiation and energy balance in boreal forests. *Journal of Geophysical Research: Atmospheres*, 108(D16).

Yatheendradas, S., Lidard, C. D. P., Koren, V., Cosgrove, B. A., De Goncalves, L. G. G., Smith, M., Geiger, J., Cui, Z., Borak, J., Kumar, S. V., Toll, D. L., Riggs, G., and Mizukami, N. (2012). Distributed assimilation of satellite-based snow extent for improving simulated streamflow in mountainous, dense forests: An example over the dmip2 western basins. *Water Resources Research*, 48(9).

Yeh, W. W.-G. (1985). Reservoir management and operations models: A state-of-the-art review. *Water Resources Research*, 21(12):1797–1818.

Yuan, X., Wood, E. F., Luo, L., and Pan, M. (2011). A first look at climate forecast system version 2 (cfsv2) for hydrological seasonal prediction. *Geophysical Research Letters*, 38(13).

Zambon, R. C., Barros, M. T. L., Barbosa, P. S. F., Francato, A. L., Lopes, J. E. G., and Yeh, W. W.-G. (2012a). *Two-Stage Stochastic Optimization of Large-Scale Hydrothermal System*, pages 2472–2481.

Zambon, R. C., Barros, M. T. L., Lopes, J. E. G., Barbosa, P. S. F., Francato, A. L., and Yeh, W. W.-G. (2012b). Optimization of large-scale hydrothermal system operation. *Journal of Water Resources Planning and Management*, 138(2):135–143.

Zeng, Z., Hsieh, W. W., Shabbar, A., and Burrows, W. R. (2011). Seasonal prediction of winter extreme precipitation over canada by support vector regression. *Hydrology and Earth System Sciences*, 15(1):65–74.

Zhang, K., Kimball, J. S., Nemani, R. R., and Running, S. W. (2010). A continuous satellite-derived global record of land surface evapotranspiration from 1983 to 2006. *Water Resources Research*, 46(9):n/a–n/a. W09522.

Zhang, L., Hickel, K., Dawes, W. R., Chiew, F. H. S., Western, A. W., and Briggs, P. R. (2004). A rational function approach for estimating mean annual evapotranspiration. *Water Resources Research*, 40(2):n/a–n/a. W02502.

Zhang, L., Potter, N., Hickel, K., Zhang, Y., and Shao, Q. (2008). Water balance modeling over variable time scales based on the budyko framework–model development and testing. *Journal of Hydrology*, 360(1):117–131.

Zhou, G., Wei, X., Chen, X., Zhou, P., Liu, X., Xiao, Y., Sun, G., Scott, D. F., Zhou, S., Han, L., et al. (2015). Global pattern for the effect of climate and land cover on water yield. *Nature communications*, 6.

Zink, M., Mai, J., Cuntz, M., and Samaniego, L. (2018). Conditioning a hydrologic model using patterns of remotely sensed land surface temperature. *Water Resources Research*, 54.