Prediction and forecasting of Evapotranspiration and Groundwater anomalies, along with improved parameterization of ET in agricultural lands

By

Ammara Talib

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Civil & Environmental Engineering)

at the

UNIVERSITY OF WISCONSIN-MADISON

2023

Date of final oral examination: 05/10/2023

The dissertation is approved by the following members of the Final Oral Committee:

Ankur R. Desai, Professor, Dept of Atmospheric and Oceanic Sciences
Paul Block, Associate Professor, Dept. of Civil & Environmental Engineering
Daniel Wright, Associate Professor, Dept. of Civil & Environmental Engineering
Jingyi Huang, Assistant Professor, Dept. of Soil Sciences

Abstract

Prediction and forecasting of Evapotranspiration and Groundwater anomalies, along with improved parameterization of ET in agricultural lands

Predicting and forecasting evapotranspiration (ET) and groundwater (GW) variations are essential for sustainable water use in agriculturally intensive areas. Despite its importance in linking energy cycles and water, ET is challenging to measure. Further, to accurately estimate ET and GW dynamics, input uncertainty and deficiencies in hydrologic models pose fundamental challenges. Moreover, for land surface model-based ET, process models GW reanalysis, and remote sensing products, performance varies with the spatiotemporal scale due to the complex nonlinear relationships among meteorological and biophysical predictors of ET and GW dynamics in managed landscapes. Because of complicated boundary conditions, heterogeneous hydrogeological characteristics, groundwater extraction, and nonlinear interactions between these factors, it is proved difficult to predict and forecast ET and GW anomalies over the long term in agricultural areas. Nevertheless, data-driven methods and deep learning have shown promising results when identifying variables' dependencies.

Here, this dissertation addresses this gap by **1)** evaluating sources of bias in the regional Wisconsin Irrigation and Scheduling Program (WISP) models and developing a correction based on eddy covariance (EC) observations. **2)** developing and evaluating the performance of data-driven models such as random forests (RF) and long short-term memory (LSTM) to predict and

forecast daily ET on diverse agricultural farms in the Midwest, USA. and **3)** utilizing recurrent neural network of LSTM as a method for forecasting GW anomalies two months in advance and for analysis of drivers that affect GW dynamics.

To accomplish the **first objective** of this dissertation, ET, observations were made for five years in agricultural fields in the Wisconsin Central Sands (WCS) region, one of the most productive agricultural regions of the country, using EC systems. WISP model ET bias was traced to the underestimation of net longwave radiation (LW_{net}) owing to a biased specification of effective clear sky atmospheric emissivity ($\varepsilon_{a,clr}$). Correcting the $\varepsilon_{a,clr}$ reduced the WISP model's bias and error for both LW_{net} and ET. The **second objective** was to apply the ET modeling framework developed for Midwest for nineteen fields where eddy covariance ET and meteorological observations are available during the growing season (April-October). In terms of daily ET prediction, a 16 parameter random forest approach outperformed a process-based land surface model. In irrigated crops, vapor pressure and crop coefficients were the most important predictors, while in non-irrigated crops, short wave radiation and enhanced vegetation index were the most important predictors. Finally, for the **third objective**, a modeling approach to forecast GW anomalies was developed and evaluated in the WCS region in the U.S. Groundwater anomalies showed high spatiotemporal variability, and their responses differed across locations depending on boundary conditions, catchment geology, climate, and topography. Land use change and irrigation pumping have interactive effects on GW anomalies forecasting.

By understanding the critical processes underlying hydrologic and climatic variability and change over land, these findings may enable improved and more accurate hydrologic and climatic

simulations. Using our framework, we can model water cycle components and dynamics in areas with unknown or uncertain subsurface properties.

Acknowledgment

I am deeply thankful to all those who have supported me throughout this journey and contributed to completing my Ph.D. dissertation. Foremost, I want to express my profound gratitude to my supervisor, Dr. Ankur Desai, whose guidance, expertise, and unwavering encouragement have been pivotal in shaping this research. His patience and valuable feedback have aided me in refining my ideas and navigating the challenges of this demanding endeavor. Working under his mentorship has been a tremendous opportunity for which I am truly grateful. Moreover, I extend my sincerest appreciation to Dr. Desai for exemplifying, both personally and professionally, that a remarkable scientist can also be an exceptional human being, demonstrating integrity and kindness. The kindness and encouragement from Dr. Desai have shaped me and provided a sense of curiosity and grounding that will forever remain a cherished part of my life's journey.

I am profoundly grateful for my family's unwavering love, understanding, and encouragement throughout the challenges and triumphs of this demanding journey. Their constant support has been a wellspring of strength and motivation that sustained me through difficult times. My brother, Abdul Sami, has been my oldest friend, and I am deeply thankful for his constant outpouring of positivity and support. I am particularly grateful to my parents, Salima Talib (mother) and Talib Ali (father), for their patience, understanding, and kindness as I navigated my scientific path. Their curiosity about my research, especially regarding groundwater and evapotranspiration interactions, forecasting, and machine learning, helped clarify my

understanding. Their warmth, words of wisdom, and unwavering support have been a guiding light. My closest friends, Saira Jamil, Asia, and Mehwish, have been pillars of support during the uncertainties and difficulties that come with pursuing a Ph.D. Their effervescent optimism and camaraderie have been invaluable on this shared journey. I am also grateful for the love and support of my extended family, including cousins, aunts, and uncles, whose presence has added immeasurable happiness to my life over the years. Furthermore, I'd also like to express my gratitude to my friends Arthur, Tom Lueptow, Barbara, Luisa, Serah, Suboth, Ted, Bereket, Cecilia, Vincent, Randy, Anusha as well as Wulan, for ensuring my well-being during my time in the United States and for their significant support and belief in me.

Since I began college, I have been fortunate to have collegiate mentors who have instilled in me the confidence to pursue engineering and the sciences, persist in the face of challenges, and ultimately apply and be accepted to graduate school. Sir Afzal, Dr. Yumna Sadef, Dr. Timothy Randhir, Dr. Jack Finn, and numerous other mentors, including those from earlier years, generously dedicated their time to review my application essays, write recommendation letters, and offer constructive criticism that helped me grow as a scientist, colleague, and individual. I am indebted to the Department of Oceanic & atmospheric sciences, Department of Soil Sciences, and Civil and Environmental Engineering faculty members at the University of Wisconsin Madison for their valuable contributions to my academic and intellectual growth. I especially want to thank my dissertation committee members, Dr. Paul Block, Dr. Daniel Wright, and Dr. Jingyi Huang. The stimulating discussions, seminars, and workshops provided an enriching environment that nurtured my research skills and broadened my understanding of the subject matter.

Besides my supervisors, I owe a great deal of gratitude to my exceptional lab mates, whose unwavering support has been a constant source of motivation throughout my journey. Over time, we evolved from mere lab partners to becoming close friends. Working with my lab mates, Ryan, Molly, and Xe, has been great friends and supporters. Ryan and Molly, thank you for listening to me as I navigated through classes and research. Thank you, Xu, for providing invaluable support while I talked about preparing for conferences. The Desai Lab has always felt like home to me since day one, and I owe this sense of belonging to my lab mates and their genuine camaraderie. Sreenath has been a true blessing, and I am thankful for his kindness, assistance, and constant support throughout my thesis writing journey. Bailey has always been there, willing to share in the excitement of my research and personal life. Andi's discipline, graceful, kind, and wise nature always made time for my inquiries, showing me the importance of prioritizing what matters. I want to thank Susi, Jess, James, Jonathan, Nikaan, and Emily for their friendship and camaraderie, which enriched the lab environment.

I am immensely thankful to my friends and colleagues from various fields who made my writing accessible to a broader audience, showed genuine concern for my well-being, and offered unwavering support during both triumphs and challenges. Moreover, I extend my deep gratitude to the anonymous readers who read, cited, commented, and shared my work online. Their quiet support reassured me that my research had a broader impact beyond academia, potentially inspiring the general public, which aligns with the main objective of my work.

To my readers, supervisors, and lab mates, I sincerely express my gratitude for dedicating their time, expertise, and constructive critiques that significantly elevated the quality of this

dissertation. Their insightful suggestions and probing questions motivated me to delve further into my research, resulting in a more robust and comprehensive body of work.

In addition to my supervisors, I am indebted to my exceptional lab mates, whose support has been a constant source of motivation. Our collaborative writing sessions and informal chats, whether conducted via screens during lockdowns or in person whenever circumstances allowed, provided a lifeline during the most challenging times. I am proud to say that we became more than just lab partners, but good friends. I was extremely fortunate to work with my labmates. I am thankful to Ryan and Molly for listening to me while I talked about navigating classes and research. Xe has been supportive for my AGU presentations. I've always felt at home at the Desai Lab, where I always felt like I belong here, from my first day. This would not have been possible without my lab mates being who they are and I'd like to thank all of them. Sreenath for being the best office mate one could ever imagine. Thank youyou're your kindness, and thank you for helping me with research and many other aspects of life and being a constant support while I was writing my thesis. Bailey, for always being there and willing to share the chaos and excitement of my research and personal life. I am grateful to have shared my office space with you. Andi, for his graceful, kind and wise soul and always making time for any of my queries and showing me that one will always have time for things should one choose so. I'd like to thank Susi, Jonathan, Jess, James, Nikaan and Emily for their friendship and camaraderie. My friends and colleagues ensured that my writing made sense to those outside of my hyper-specific field, checked in on my physical and mental health, and were there for me whether I succeeded or failed. Similarly, I must thank the anonymous readers who have read my work online, cited it, commented on it, or shared it

with others, for their silent support. Their readership assured me that my research was reaching people outside of my small, academic circle and potentially inspiring the greater public which I strive to serve. I would like to thank them for their time, expertise, and constructive critiques that significantly improved the quality of this work. Your thoughtful suggestions and probing questions pushed me to delve deeper into my research, resulting in a more robust and comprehensive dissertation.

I am also grateful to the support staff and administrative personnel of Department of Oceanic & atmospheric sciences and Civil and Environmental Engineering, especially Sue Foldey, Cheryl Loschko, Christi Levenson, Stacey Koch for their assistance and efficiency in handling various administrative matters, allowing me to focus on my research without unnecessary distractions. This work was financially supported through several research grants and by my advisor. I acknowledge support from the Wisconsin Vegetable and Potato Growers Association award to UW-Madison, the Wisconsin Department of Natural Resources, and the UW Center for Climatic Research Climate, People, and Environment Program, and thank Jonathan Thom, T Houlihan, J Pavelski for site support at US-CS1 and US-CS3-US-CS6. DR and JC are supported by the NASA Carbon Cycle & Ecosystems program (NNX17AE16G). I also acknowledge support from the Wisconsin Groundwater research and monitoring program, the Wisconsin Department of Natural Resources, and the UW Center for Climatic Research Climate, People, and Environment Program. Among the most memorable highlights of my PhD journey was the opportunity to participate in the fieldwork for setting up towers at US-CS1 and US-CS3. This enriching experience would not have been possible without the dedication and expertise of Dr. Desai and Jonathan

Tom. I am deeply grateful for the knowledge, connections, and friendships that were forged during that time.

Lastly, I want to express my gratitude to all the participants who took part in the studies that contributed to this dissertation. Your willingness to share your time and experiences has been crucial in generating the empirical evidence that underpins this research. To everyone mentioned above and to those whose names may not appear here but have contributed in various ways, thank you from the bottom of my heart. This achievement would not have been possible without your collective support and belief in me.

Contents

Abstract i

Acknowledgment iv

1. Background and motivation 1

1.1 Evapotranspiration and ground water; vital components of the water cycle 1

1.2 Existing approaches to measuring ET and GW and their limitations 4

1.3 Data-driven or machine learning approaches to estimating ET and GW 6

1.4 Research questions and Scope 10

2. Improving parameterization of a regional irrigation scheduling program for evapotranspiration estimation with eddy covariance measurements 13

Abstract 13

Key Points 14

2.1 Introduction 14

2.2 Methods 19

2.2.1 Site description 20

2.2.2 Wisconsin Irrigation, scheduling, and planning (WISP) Model 22

2.2.3 Inputs to the WISP Model 25

2.2.4 Meteorology and Eddy covariance observations 26

2.2.5 Approaches to evaluation and correction of WISP model 28

2.2.6 Model evaluations 30

2.3 Results 30

2.3.1 ET and meteorological observations comparison to WISP 30

2.3.2 LW_{net} comparison before and after correction of WISP model 31

2.3.3 WISP model evaluation for ET_a comparison before and after correction 34

2.3.4 WISP model ET_a evaluation by month 45

2.4 Discussion 48

2.4.1 Role of emissivity 48

2.4.2 Corrected model performance by site and month 52

2.4.3 Model limitations and potential improvement 54

2.5 Conclusion 56

2.6 Acknowledgements 57
2.7 Bibliography 58
3. Evaluation of prediction and forecasting models for evapotranspiration of agricultural lands in the Midwest U.S. 70
Abstract 70
Key Points 71
3.1 Introduction 72
3.2 Methods 77
3.2.1 Data description 78
3.2.2 Random Forest model framework 80
3.2.3 Long Short-Term Memory Network (LSTM) 81
3.2.4 Land Surface Model 84
3.2.5 Significant predictors 85
3.2.6 Forecast model 86
3.2.7 Model evaluations 87
3.3 Results 88
3.3.1 RF versus LSTM prediction model evaluation 88
3.3.2 Significant predictors 90
3.3.3 Model performances 91
3.3.4 Forecast model results 93
3.4 Discussion 95
3.4.1 Model evaluations 95
3.4.2 Significant predictors 100
3.4.3 Forecast models evaluations 103
3.4.4 Limitations and future directions 105
3.5 Conclusion 107
3.6 Acknowledgements 108
3.7 Bibliography 127
3.8 Supplemental Information 139
Chapter 4 145
4. Spatial and Temporal Forecasting of Groundwater anomalies 145
Abstract 145

Key points 146
4.1 Introduction 147
4.2 Methods 152
4.2.1 Description of study area 152
4.2.2 Input data: 154
4.4.3 Model Development 159
4.4.4 Models to study the influence of drivers on GW anomalies forecast. 163
4.3 Results 164
4.3.1 Temporal model performance for two-month GW anomalies forecast 164
4.3.2 Spatial model performance for two-month GW anomalies forecast 169
4.3.3 Comparison of temporal and spatial forecast model 173
4.3.4 Effect of drivers on GW anomalies forecast 174
4.3.5: Land use change and GW pumping effect 175
4.4 Discussion 179
4.4.1 GW anomalies forecast models performance 179
4.4.2 Model forecasts and hydrometeorological characteristics 181
4.4.3 Relations between simulation results and catchment attributes 182
4.4.4 Land use change and irrigation pumping effect 183
4.5 Conclusion 186
4.6 Acknowledgment 189
4.7 Bibliography 189
5. Conclusion 198
5.1 Summary & Outlook 198
5.2 Improving parameterization of a regional irrigation scheduling program for evapotranspiration estimation with eddy covariance measurements 200
5.3 Evaluation of prediction and forecasting models for evapotranspiration of agricultural lands in the Midwest U.S. 201
5.4 Spatial and Temporal Forecasting of Groundwater anomalies 202
5.5 Limitations & Future work 204

Chapter 1

1. Background and motivation

1.1 Evapotranspiration and ground water; vital components of the water cycle

The continents' major food production regions have also been affected by extreme weather patterns. US agriculture was adversely affected in 2011 by a severe drought affecting 80% of agriculture (AghaKouchak et al., 2014). These changes are therefore putting greater stress on resources and crops, and increased irrigation water demand worsens global water scarcity. Water stress and crop productivity need to be alleviated by irrigation expansion due to global warming and population growth (Zhang 2020; McDermid et 2021). When groundwater is extracted for irrigation of crops, it can result in a prolonged period of water depletion (Foley, 2011). Water is extracted from thousands of high-capacity wells to irrigate crops in Wisconsin Central Sands, a highly productive area in the United States (Luczaj and Masarik, 2015; Nocco et al., 2018). Due to recurring wet and dry periods, irrigation-induced stress continues even during normal to moderately wet periods in Central Sands (Haucke et al., 2016). The impact of groundwater withdrawal on surface waters and the risk of future climate extremes raises questions about its long-term sustainability.

Understanding terrestrial ecosystem processes in a changing climate, such as flash droughts (Kim et al., 2019; Otkin et al., 2016), water resource management (e.g., irrigation efficiency), requires accurate estimation and forecasting of ET and GW anomalies (Shugart, 1998; Anderson et al., 2011). Prediction and forecasting of ET and GW fluctuations is a fundamental

component of hydrological applications aimed at conserving water resources, especially for irrigation. To manage water resource challenges in sustainable agriculture, a tool to predict and forecast ET and GW anomalies can be useful (Perera et al., 2014; Moratiel et al., 2020). Furthermore, the accurate measurement of irrigation water losses to the atmosphere through ET is necessary to design effective water policies and efficiently manage irrigation water.

As temperatures increase, the water cycle undergoes more extreme and variable changes. Both theoretical studies and climate models anticipate higher evapotranspiration (ET) levels (Syed et al., 2010; Famiglietti, 2014; Fisher et al., 2017; Vahmani et al., 2022). ET represents a significant part of global land surface water budgets (Wang and Dickinson, 2012; Barr et al., 2014), and it forms a connection between terrestrial water and the atmosphere (Priestley and Taylor, 1972; Donohue et al., 2010; Wei et al., 2017). In addition to ET, groundwater (GW) also plays a crucial role in the water cycle, providing about 30% of freshwater reserves (Schneider et al., 2011) and serving as a major source of drinking water worldwide (Luyssaert et al., 2014; Majumdar et al., 2020; Al-Yaari et al., 2022). However, expanding irrigated crop fields and increasing irrigation practices have led to disruptions in the groundwater system, exacerbating hydrologic extremes such as floods and droughts (Milly et al., 2002).

Irrigation significantly contributes to freshwater withdrawals, using approximately 70% of total water withdrawals from surface and groundwater globally (Grafton et al., 2017; Koch, 2020). This irrigation expansion affects not only agricultural productivity but also local and regional atmospheric processes, precipitation patterns, and climate (Pokhrel et al., 2012; Lu et al., 2017; Hurtt et al., 2020). Moreover, extreme weather patterns have impacted major food production

regions, with severe droughts causing adverse effects on US agriculture in 2011 (AghaKouchak et al., 2014). The combination of global warming, population growth, and increased irrigation demand is intensifying water stress and jeopardizing crop productivity, necessitating irrigation expansion to alleviate these challenges (Zhang 2020; McDermid et 2021). However, excessive groundwater extraction for irrigation can lead to prolonged water depletion, raising concerns about long-term sustainability (Foley, 2011; Luczaj and Masarik, 2015; Nocco et al., 2018; Haucke et al., 2016).

Given these challenges, accurately estimating and forecasting ET and GW anomalies becomes crucial for managing water resources effectively, especially for sustainable agriculture (Shugart, 1998; Anderson et al., 2011). Accurate measurement of irrigation water losses through ET is necessary for designing effective water policies and optimizing irrigation practices (Perera et al., 2014; Moratiel et al., 2020). Understanding terrestrial ecosystem processes in a changing climate, such as flash droughts, further emphasizes the need for reliable prediction and forecasting of ET and GW fluctuations (Kim et al., 2019; Otkin et al., 2016). Ultimately, having a tool to predict and forecast ET and GW anomalies can help address water resource challenges and promote sustainable water management in agriculture.

1.2 Existing approaches to measuring ET and GW and their limitations

ET, the second largest component in the land surface water budget, is challenging to measure at large scales, resulting in limited observational constraints (Teuling et al., 2009; Bohn & Vivoni, 2015). Hydrologic studies often focus on the supply side, neglecting the demand side, particularly the impact of ET, which has become increasingly critical due to climate-driven and management-driven water demands and droughts. Various methods, including direct point observations and remote sensing, can estimate ET_a and PET at fine spatiotemporal scales (Talib et al., 2021; Jung et al., 2010; Sörensson & Ruscica, 2018). Direct measurements, such as those from eddy covariance towers, can accurately assess ET, but regional and long-term studies face challenges in terms of cost, logistics, and measurement scale. To predict and forecast ET for broader applications, readily available drivers are required.

Researchers have also combined data assimilation methods and land surface models to estimate ET, but the estimates still have a relatively high error range due to various factors (Zou et al., 2017; Xu et al., 2018; Vinukollu et al., 2012; Lian et al., 2018; Long et al., 2014). Few studies focus on forecasting actual ET in irrigated and non-irrigated areas, even though many estimate or predict reference ET under different climatic conditions e.g., (Fang et al., 2018, Kimball et al., 2019). To better understand hydrological and biogeochemical processes and monitor groundwater dynamics, direct measurements and modeling studies have been conducted to analyze the human water footprint (Wada et al., 2017). However, simulating groundwater levels is challenging due to the complexity of groundwater systems, heterogeneity, and time-lag

dependencies (Clark et al., 2015). Hydrological variables and system boundary conditions also significantly affect groundwater dynamics, such as climate variability and pumping (Hintze et al., 2020; Cai et al., 2021).

Satellite data, employing remote sensing techniques, allows monitoring large areas, and numerous space-borne missions provide products for assessing various quantities related to the global water cycle. For example, MODIS (Moderate Resolution Imaging Spectroradiometer) has been used to quantify terrestrial Earth surface radiation (ET), GPM (Global Precipitation Measurement) has been used to quantify precipitation, SMAP (Soil Moisture Active Passive) has been used to quantify soil moisture, and USDA-NASS has been used to quantify land use (MardanDoost et al., 2019). Specifically, models based on remote sensing products; for example, the two-source energy balance (TSEB), the Operational Simplified Surface Energy Balance (SSEBop Mecikalski et al., 1999), Simplified Surface Energy Balance (SSEB, Senay et al., 2007**)**, the Priestly-Taylor (PT) based JPL retrieval algorithm (L3_ET_PT-JPL Fisher et al., 2008**),** hydrological models (Cai et al., 2009; Wanders et al., 2014; Kittel et al., 2017; Fisher et al., 2020; Zhang & Long, 2021), models based on vegetation index (Glenn et al., 2010), or other statistical and empirical models (Wang & Liang, 2008; Chen et al., 2021) have also been implemented. Although remote sensing estimates are increasingly used to estimate ET and GW dynamics in irrigated agricultural areas, the resolution of these products is too coarse for local estimates of ET and GW flux, and they haven't been related to forecasts of spatial region GW anomalies and drivers.

1.3 Data-driven or machine learning approaches to estimating ET and GW

Using machine learning (ML) techniques, it becomes feasible to merge data from easily available predictors obtained through remote sensing and ground observations to forecast ET without relying on field-based physical parameters. Numerous data-driven models have been employed in ET simulation studies, such as those conducted by Deo and Ahin (2015) and Fang et al. (2018). ML algorithms enable us to estimate and forecast future scenarios by capturing nonlinear relationships in time series or spatial data. For instance, Yang et al. (2006) and Tabari et al. (2012) employed support vector machine (SVM) methods to estimate an eight-day average ET and reference ET using ground observations and remote sensing. Landeras et al. (2009) utilized a regression tree approach to forecast weekly reference ET, while Bodesheim et al. (2018) employed a random forest approach to estimate ET. Random forest (RF) is particularly noteworthy for handling high-dimensional regression problems and identifying interactions between model predictors without explicit training (Auret and Aldrich, 2012; te Beest et al., 2017). Research by Shiri et al. (2018) suggests that tree-based models can provide more accurate estimates of reference ET compared to wavelet-random forest models. Additionally, this approach is flexible, simple, robust, and avoids overfitting, utilizing ensemble trees and randomization to achieve reliable performance on both training and test data (Chen et al., 2020). Apart from ensemble tree algorithms, artificial neural networks (ANNs) have also been applied for ET prediction in studies by Abdullah et al. (2015), Pandey et al. (2017), and Ferreira et al. (2019).

Machine learning-based hydrological models have proven effective in simulating groundwater levels, as demonstrated by studies conducted by Yoon et al. (2011) and Malekzadeh

et al. (2019). Artificial neural networks (ANNs) are a type of machine learning model that has been successfully utilized to predict nonlinear hydrological processes, including precipitation, runoff, stream flow, and water quality. Although these models do not explicitly consider the physical meanings of the variables during the machine-learning process, their simulation results can often be interpreted with multiple physical explanations (Brodrick et al., 2019). Catchment characteristics play a significant role in hydrology-related variations. Several previous studies have examined the primary recharge methods for groundwater levels in the continental United States (CONUS) based on regional characteristics (Gleeson et al., 2012), suggesting that machine learning techniques can be employed to analyze drainage basin characteristics and their impact on groundwater fluctuations.

Several approaches involving artificial neural networks (ANNs), such as the convoluted neural network (CNN), have been developed for modeling ET and GW levels (Tavares et al., 2015; Yassin et al., 2016). However, conventional feed-forward ANNs are not well-suited for time series analysis as they lack information about input sequences and require complex data pre-processing, such as singular spectrum analysis, to address this limitation (Sahoo et al., 2017). Additionally, traditional ANNs are prone to the problem of exploding gradients (Rangapuram et al., 2018). To overcome these issues and understand temporal dynamics in sequential data, recurrent neural networks (RNNs) have emerged as a suitable neural network architecture (Carriere et al., 1996).

In the field of RNNs, several examples include convolutional neural networks, long short-term memory (LSTM), and genetic programming (Kratzert et al., 2019). LSTM, a specific type of RNN, effectively solves the problem of vanishing gradients, making it ideal for applications like ET

and GW anomalies prediction and time series forecasting, where the order of input variables is crucial. For our study, we have chosen to utilize RNNs, specifically LSTM, in combination with ensemble tree algorithms like random forest (RF) because our focus is on time series prediction and forecasting.

LSTM, through its connections between units and cells, facilitates data movement both forward and backward, effectively eliminating lagged dependencies present in traditional RNNs. By employing this approach in the context of the water cycle, it becomes possible to retain past information for future applications, such as accounting for water storage effects like snow or shallow groundwater-driven systems. Kao et al. (2020) demonstrated the utility of LSTM models in forecasting floods in flood-prone regions by linking rainfall sequences with runoff sequences. Moreover, in Kratzert et al.'s (2019) study, LSTM outperformed benchmark physically-based coupled models in runoff simulation for various watersheds, demonstrating its effectiveness based on process-based constraints. Despite hydrologists utilizing ML for addressing various hydrologic problems for many years, the field is still in its early stages, with numerous long-standing and unconventional data sources remaining largely unexplored and fragmented. Moreover, the presence of incomplete and diverse information continues to be a significant challenge (Bergen et al., 2019).

Existing methods for predicting and forecasting actual ET face various challenges, including the need for extensive parametrization, the lack of relevant data drivers, the computational complexity of process-based models, and the inability to directly estimate actual

ET and GW anomalies using empirical models. Additionally, there is a knowledge gap regarding the performance of data-driven models under different soil types and for various types of irrigated and non-irrigated crops. To address these challenges and improve hydrological predictions in ungauged catchments, the integration of gridded ET and hydrometeorological datasets with ground observations and machine learning models shows promise (Huang et al., 2020). For instance, eddy covariance observation data have been widely utilized to estimate uncertainty in ET products. The inclusion of EC observation sites is crucial for capturing heterogeneous ET through remote sensing and point observations, considering the influences of topography, soils, and vegetation (Baldocchi et al., 2001; Sriwongsitanon et al., 2020; Li et al., 2019; Talib et al., 2021). Moreover, combining data that describe these processes across multiple sectors can complement the modeling efforts. Machine learning (ML) is well-suited for handling nonlinear and nonparametric relationships empirically. However, current studies applying ML models have been limited to a small number of test datasets. In the context of agricultural fields, few studies have explored the relative contributions of different input datasets (predictors) to ET and GW anomalies models, especially with respect to different management methods (irrigated vs. rainfed), crop types, and soil texture. Furthermore, despite research on various components of the water cycle, remote sensing datasets are rarely combined to forecast spatial GW anomalies and evaluate drivers, which hinders the estimation of these fluxes under specific conditions.

1.4 Research questions and Scope

The Wisconsin Irrigation and Scheduling Program (WISP) is often used by farmers in the Central Sands Region of Wisconsin (WCS) to schedule irrigation (https://wisp.cals.wisc.edu/). Using the Priestley-Taylor (PT) approach, WISP provides PET estimates that can be scaled to ET based on the percent canopy cover. It is important to note that even though ECs and micrometeorological observation stations have been increasingly used to calibrate and validate coarser-scale global data sources (Fisher et al., 2011; Fisher et al., 2020), the WISP model has not yet been extensively evaluated using direct EC-based ET observations, despite personal communications with local land managers indicating that the WISP model produces unreliable ET estimates for irrigation management. *Moreover, we seek to broaden the scope of our research to understand ET dynamics in the Midwest in this dissertation. Furthermore, we seek to assess GW dynamics along with ET. It is because severe irrigation pumping impacts have been observed in this region during relatively dry weather.* We will gain a better understanding of aquifers' physical mechanism, natural variability, and response through this dissertation. Both Et and GW dynamics prediction and forecast data sets would be useful for water managers and food security efforts in delineating agricultural areas that are under stress, revealing factors that influence GW fluctuation behavior, and clarifying how climate change and biophysical parameters affect GW dynamics.

The overall objective of this dissertation is to apply new observation and modeling tools to better understand and forecast the dynamics of ET and GW in agricultural regions, especially those facing water stress from drought and increasing water withdrawals.

Research Objective 1

One objective of this dissertation is to assess uncertainties in the WISP model inputs due to meteorological variables and improve model performance by calibrating parameters based on ground EC observations and satellite measurements.

The novelty of this study is that it quantifies the sources of errors of a regional-scale ET model in irrigated agricultural fields and develops a simple and scalable correction approach for improvement of existing ET monitoring platforms for agricultural water management.

Research Objective 2

Here, we also ask how well empirical ML models can predict and forecast ET 3 days in advance in irrigated and rain-fed agricultural lands across the Midwest US and what are important drivers for predicting and forecasting ET 3 days in advance in irrigated and non-irrigated areas.

The results of this evaluation allow us to better understand the predictors of accuracy and uncertainty in the ET models and propose a multistep prediction and forecast agricultural ET model that can be applied to locations with limited *in situ* observations.

Research Objective 3

We ask how can easy-to-obtain remote sensing derived environmental datasets be combined to accurately forecast GW anomalies across densely irrigated regions, using computationally efficient ML models, how do GW anomalies forecast vary in space and across the WCS, and how irrigation pumping, and crop type change can influence this anomaly forecasting.

It would be helpful for water managers and food security efforts to have spatial monthly GW anomalies forecast data sets that accurately delineate agricultural areas under stress, give insight into factors influencing GW fluctuation behavior, and clarify how climate change and biophysical parameters affect GW dynamics.

In this dissertation, prediction refers to a model estimate of output at timestep based on model predictors from time step t. Forecasting refers to model output for a specific future point (e.g., t+1, t+2) based on model predictors from time step t as well as predictors from t+1 and t+2 (e.g., ET forecast three days in advance with incoming SW from t and t+1 and t+2).

Chapter 2

2.[1]Improving parameterization of a regional irrigation scheduling program for evapotranspiration estimation with eddy covariance measurements

Abstract

Actual evapotranspiration (ET_a) is an essential variable in linking energy cycles, carbon, and water, yet challenging to measure. Input uncertainty and deficiencies in the key elements of hydrologic models are fundamental challenges for optimizing model performance. Furthermore, the performance of land surface model-based ET_a, reanalysis, and remote sensing products varies with spatiotemporal scales. Here, we evaluate sources of bias in the regional Wisconsin Irrigation and Scheduling Program (WISP) model and develop a correction using eddy covariance (EC) observations. ET_a, observations were made for five years (2018–2022) using EC systems in agricultural fields in Wisconsin. WISP model ET_a bias was linked to underestimation of net long wave radiation (LW_{net}) that was traced to incorrect specification of effective clear sky atmospheric emissivity ($\varepsilon_{a,clr}$). Applying a correction to the $\varepsilon_{a,clr}$ led to reduced WISP model percent bias (pbias) and error for both LW_{net} and ET_a. The calibrated model more accurately

[1] Chapter two has been submitted in the journal of Agricultural and Forest Meteorology for review (AGRFORMET-D-23-00723)

represented observed ET_a. The results indicate that explicit treatment of the LW_{net} balance decreases the uncertainty of model parameters and improves the WISP model performance at independent sites. Applying this improved model parameterization reduced the bias of LW_{net} radiation from 62.8% to -6.2%, which improved the Nash-Sutcliffe Efficiency (NSE) from -0.08 to 0.52 for ET_a on training sites. Additionally, overall pbias was significantly reduced ($p=0.035$) for validation sites after WISP correction. Hence, WISP performance improved for different crop types when optimal regional parameters were used, confirming the physical parameters' reliability. Our results highlight that model development should focus on energy balance parameterizations to improve ET simulation to improve the accuracy of hydrologic and climatic simulations for understanding critical processes underlying hydrologic and climatic variability and change over land.

Key Points

2 Actual evapotranspiration; Eddy Covariance; Longwave radiation; emissivity; percentage bias

2.1 Introduction

The water cycle is experiencing more extremes and greater variance, and both theory and climate modeling studies predict increased evapotranspiration (ET) due to rising temperatures (Huntington, 2006; Syed et al., 2010; Kunkel et al., 2013; Famiglietti, 2014; Fisher et al., 2017; Lauffenburger et al., 2018; Pool et al., 2021; Vahmani et al., 2022). Extreme weather patterns have also been observed across key food production regions. For example, a severe drought in the midwestern United States impacted 80% of US agriculture in 2011, followed by an even more

severe drought in 2012 (AghaKouchak et al., 2014). Consequently, resources and crops are subjected to increased stress due to these changes, and global water scarcity is exacerbated by increased irrigation water demand. Irrigation, for instance, uses 70% of total water withdrawals from surface and groundwater worldwide, which is crucial for food production (Grafton et al., 2017; Koch, 2020). A combination of global warming and population growth requires irrigation expansion to alleviate water stress and maintain crop productivity (Zhang 2020; McDermid et 2021).

A prolonged period of water depletion can result from the extraction of groundwater for the irrigation of crops (Foley, 2011; Dalin et al., 2017). For instance, irrigating crops in Wisconsin Central Sands, a highly productive area in the United States, involves the extraction of water from thousands of high-capacity wells (Kraft et al., 2012; Luczaj and Masarik, 2015; Nocco et al., 2018). Water recharge estimates in Central Sands suggest that irrigation-induced stress to groundwater table continues even during normal to moderately wet periods, and aquifer depletion can be episodic because of recurring wet and dry periods (Haucke et al., 2016). Therefore, groundwater withdrawal for irrigation comes with questions about its long-term sustainability due to its impacts on surface waters and the increasing risk of future climate extremes. A major loss term of irrigated water is evapotranspiration (ET). To design effective water policies and to manage irrigation water efficiently, it is necessary to accurately quantify losses of irrigation water to the atmosphere through ET (Sorooshian et al., 2012; Zaussinger et al., 2019).

ET is the second largest term of the terrestrial water budget and accounts for around 65% of precipitation over land (Bengtsson et al., 2010), yet it is also one of the least constrained by observations since it is difficult to measure directly at large scales (Monteith et al., 1965; Teuling et al., 2009; Bohn and Vivoni, 2015). Most hydrologic studies have also focused on the supply side (e.g., groundwater, soil moisture, precipitation, snow) while ignoring the demand side (e.g., ET) (Ohmura et al., 2002; Rosenbaum et al., 2012; Argus et al., 2017; Enzminger et al., 2019; Tennant et al., 2020). Understanding both sides of the supply-demand equation has become more crucial due to increased water demands (both climate-driven and management-driven) and droughts, particularly the loss of water through ET, when trying to mitigate vegetation and crop stress. Terrestrial ET is also critical for linking energy changes (latent heat flux), water, and carbon dioxide (transpiration-photosynthesis trade-off) (Monteith, 1965; Wang and Dickinson, 2012; Xing et al., 2016; Orth and Destouni, 2018; Althoff, 2020; de Oliveira et al., 2021; Liu et al., 2022). ET is the dominant variable necessary for agricultural water management; irrigation applies water to approximate atmospheric ET demand) and the leading indicator of extreme flash drought events. Furthermore, it is also possible to use changes in ET to diagnose climate variability and change, such as whether or not the land surface dries or wets over a decade. (*Mao et al.,* 2015; Mokhtar et al., 2020).

The estimation of ET can also be an invaluable tool for water managers who need to allocate dwindling water resources to maximize productivity and benefit society while reducing ecological, legal, social, and economic costs (Djaman et al., 2018; Schwartz et al., 2020). Since irrigation significantly increases ET in the soil-plant-atmosphere continuum, ET data sets are also

useful for estimating irrigation water use. The estimations of potential ET (PET) and actual ET (ET_a) can be carried out using different methods based on different underlying principles (Allen et al., 2011). These methods can cover fine spatiotemporal scales by direct point observations (Talib et al., 2021) or coarser scales through indirect methods employing remote sensing products (Jung et al., 2010; Sörensson and Ruscica, 2018). Remotely sensed ET observations have recently been used to estimate irrigation water use at larger scales. Specifically, models based on remote sensing products; for example, the two-source energy balance (TSEB), the Operational Simplified Surface Energy Balance (SSEBop Mecikalski et al., 1999), Simplified Surface Energy Balance (SSEB, Senay et al., 2007)**,** the Priestly-Taylor (PT) based JPL retrieval algorithm (PT-JPL Fisher et al., 2008), hydrological models (Cai et al., 2009; Wanders et al., 2014, Hain et al., 2015; Kittel et al., 2017; Fisher et al., 2020; Zhang and Long, 2021), models based on vegetation indices (Glenn et al., 2010), or other statistical and empirical models (Wang and Liang, 2008; Chen et al., 2021) have been implemented.

The development and analysis of global gridded ET datasets, coupled with remote sensing data and modeling, provide the opportunity to enhance hydrological predictions in ungauged catchments (Huang et al., 2020). The gridded ET datasets, however, have large uncertainties due to differences in parameterization schemes, model structure, and forcing data (Khan et al., 2018; Blöschl et al., 2019; Giles-Hansen, 2021). Additionally, the ET estimates used in these studies are usually of coarse spatial (1-5 km) and temporal (daily-monthly) resolutions (Anderson et al., 2012; Holmes et al., 2018; Melo et al., 2021). Furthermore, the quality of remote sensing data is not always guaranteed due to the varying accuracy of remote sensing data across regions due to

assumptions within retrieval algorithms, and challenges related to repeat intervals and cloud cover (Liu et al., 2016; Sun et al., 2018). Input data for such models should therefore be carefully selected. ET products have been evaluated at regional and global scales using various methods (Ma et al., 2021). For instance, eddy covariance (EC) observation data has been widely used for estimating uncertainty in ET products (Baldocchi et al., 2001; Joshua et al., 2020; Sriwongsitanon et al., 2020; Cawse-Nicholson et al., 2021). Furthermore, EC observation sites are often needed for both point observations and remote sensing platforms to capture heterogeneous ET (Li et al., 2019), because ET depends on meteorological forcing, topography, soil, and vegetation (Talib et al., 2021).

In the Central Sands Regions of Wisconsin (WCS), irrigation is fundamental in supporting regional agricultural production and rural community development (Nocco et al., 2018). For irrigation scheduling, farmers often rely on the Wisconsin Irrigation and Scheduling Program (WISP) model, developed by soil scientists at UW-Madison and based on satellite data (https://wisp.cals.wisc.edu/). WISP is a regional model based on the Priestley-Taylor (PT) approach and provides PET estimates that can be scaled to ET based on a canopy cover model (Henggeler et al., 2010; Sanford and Panuska, 2015). Without addressing advection effects, PT calculates potential evapotranspiration (PET) from a well-watered crop surface. A number of hydrologic models have used the PT method, for example, the Soil and Water Assessment Tool (SWAT), to represent the upper limit of ET (Arnold et al., 2012; Zeng et al., 2014). Crop models also use the PT method to estimate PET and irrigation water demands (Deryng et al., 2011; Hoogenboom et al., 2019).

While EC and micrometeorological observation stations have recently been used in greater frequency for calibration and validation of coarser-scale global data sources (Fisher et al., 2011; Fisher et al., 2020), the WISP model has not been extensively evaluated based on direct EC-based ET ground observations although personal communications with local land managers indicate the WISP model can produce unreliable ET estimations for irrigation management in some cases. Our working hypothesis is that EC and associated meteorological and radiation measurements can be used to assess the error sources and reduce the uncertainty of the model parameters to improve the performance of the WISP model. Although previous studies have sought to correct remote sensing products (Chirouze et al., 2014; Huang et al., 2020; Delogu et al., 2021; Feng et al., 2021), the sources of uncertainties vary with the model principles and input datasets. Therefore, the main aims of the present study are to 1) assess uncertainties in the WISP model inputs due to meteorological variables and 2) improve model performance by calibrating parameters based on ground EC observations and satellite measurements. The novelty of this study is that it quantifies the sources of errors of a regional-scale ET model in irrigated agricultural fields and develops a simple and scalable correction approach for the improvement of existing ET monitoring platforms for agricultural water management.

2.2 Methods

In this paper, the WISP model input comparison is made with observations to assess model input uncertainty and physical assumptions. EC observations were used to calibrate the

WISP model regionally. Then the calibrated model is validated at different sites for net longwave radiation (LW_{net}) and ET.

2.2.1 Site description

The WCS study area (Fig. 2.1) covers 630,000 acres bordered by the Wisconsin River on the west and the headwater streams of the Fox and Wolf River basins on the east. Water is typically found between 3 and 20 meters below the surface of the sandy mantle, which contains the region's aquifer. In addition to over 80 lakes, the area includes over 1000 kilometers of headwater streams and wetlands strongly connected to groundwater. As a result, they are prized for their ecosystems, which support warm-water fisheries, cold-water salmonids, related species, and rare and endangered plants (Kraft, 2012). Rainfall in the WCS typically ranges from 790 to 810 mm per year, with 60 to 70% returning to the atmosphere as ET and 30 to 40% recharge (Kucharik, 2010; Motew and Kucharik, 2013). The soil at the study site is moderately well-drained to excessively drained.

The WCS is a typical example of irrigated agriculture in the northern Great Lakes region. Aquifers that support headwater trout streams, lakes, and extensive wetlands also provide groundwater for WCS farmers via high-capacity wells (Kraft, 2012). The WCS has over 2,100 high-capacity wells that irrigate approximately 86,500 ha of potato, corn (field and sweet), snap beans, and peas, which are produced for US$550 million (Smail, 2016). Irrigation was considered economically advantageous in the WCS since it tripled yields and reduced drought risk on sandy, fast-draining soils. As irrigated lands were converted, by the mid-2000s, there were over 3000 high-capacity wells in the WCS (Weeks and Stangland, 1971). While only modestly dry conditions

prevailed, reservoir pumping caused severe impacts on streams, lakes, and wetlands (Kraft, 2010). As a result, stretches of the Little Plover River, a Class I trout stream with recreational value for aquatic stakeholders in the WCS, were dried out seasonally from 2005 to 2009 (Bradbury et al., 2017). Water scarcity, surface water stress, and ongoing irrigation expansion in this region are causing conflict over water resources.

A center-pivot irrigation system pumps and distributes groundwater from high-capacity wells, altering evapotranspiration and recharge in these landscapes, which alters water supply to streams, lakes, and wetlands. During the last 60 years, studies have documented a significant increase in precipitation under native potential vegetation of 100 to 200 mm y^{-1} and 50 to 100 millimeters from June to August (Motew and Kucharik, 2013). Some 70% of this precipitation was partitioned to groundwater recharge. In the absence of conversion to irrigated land cover, groundwater recharge should have increased, dispelling the notion that climate change is responsible for groundwater declines and surface water stresses in the WCS. Water level declines and hydrological stresses across the WCS region can be explained by a 45 mm reduction in annual recharge (equivalent to net consumed irrigation) on irrigated lands (Kraft, 2012). Since the WCS has flat, sandy terrain that has negligible overland flow, recharge is modeled as the difference between precipitation and ET, where changes in soil moisture storage are incorporated at weekly or shorter timescales and assumed negligible at monthly and longer timescales due to the limited capacity for holding water (Weeks et al., 1965; Tanner and Gardner, 1974). Due to the high infiltration rates of sandy soils, surface runoff is also considered negligible at the field scale. Weighing lysimeters and crop coefficient approaches have been used to model

evapotranspiration for different WCS vegetable crops and other land covers (Black et al., 1970, Nocco et al., 2017). Using the assumption that precipitation stays the same regardless of crop type, differences in recharge have been measured and modeled as differences in ET (Kraft et al., 2012). Water budget estimates at national and regional scales indicate that vegetation type and phenological stage are the most critical determinants of recharge and ET in irrigated agroecosystems.

2.2.2 Wisconsin Irrigation, scheduling, and planning (WISP) Model

The WISP model uses the Priestly-Taylor model to estimate PET. The Priestly-Taylor model (Priestly and Taylor, 1972) is shown in Equation (1) and is based on the assumption that the equilibrium term $\frac{\Delta}{\Delta+\gamma}(R_n - G)$ is significantly larger than the aerodynamic term. Where Δ is the slope of saturation vapor pressure-Temperature curve, λ is the psychrometric constant, R_n is net radiation (Wm^{-2}), G is soil heat flux (Wm^{-2}), λ_v is the volumetric latent heat of vaporization, 2453 MJ m^{-3} and α is the Priestly- Taylor coefficient for which we used the previously adjusted (Sanford and Panuska, 2015) value of 1.28 for the study area. The net shortwave radiation (R_n) resulting from the balance between incoming and reflected solar radiation is given by:

$$R_n = (1 - \alpha albedo)\, R_s - LW_{net} \qquad (2)$$

Where R_s is incident shortwave radiation (MJ m^{-2} day^{-1}) and albedo (or canopy reflection coefficient) was assumed to be 0.25 (dimensionless) for the crops. The Stefan–Boltzmann equation was used to estimate the longwave radiation emitted from the atmosphere to the surface.

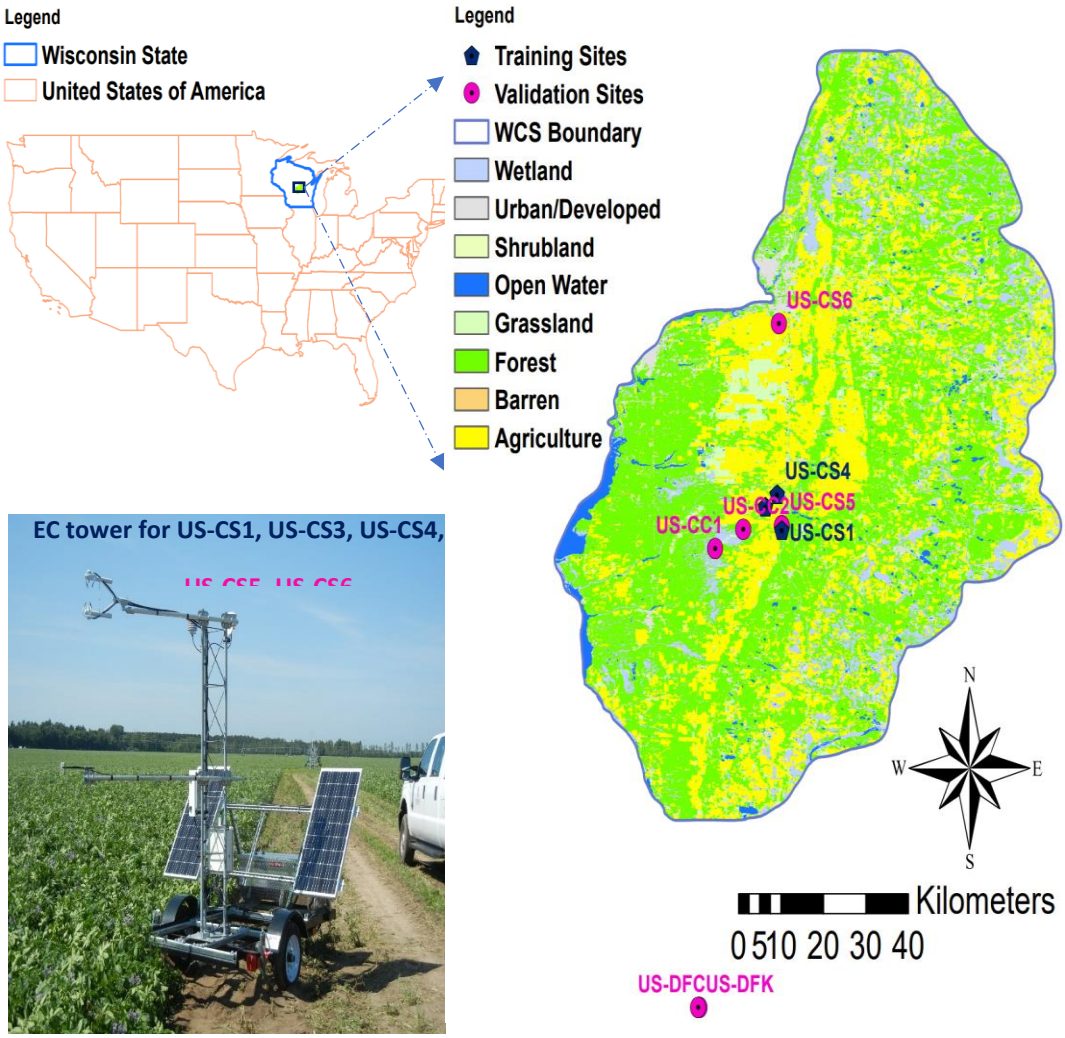

Figure 2.1: Study site is located in Wisconsin, USA. Land use is shown along with locations of training and validation sites. Portable EC tower used for measurements on US-CS1, US-CS3, US-CS4, US-CS5, US-CS6 is also presented.

$$PET = \alpha \,.\frac{\Delta}{\Delta+\gamma}(R_n - G) \qquad (1)$$

The longwave radiation emitted from the atmosphere was estimated by the Stefan–Boltzmann equation:

$$LW_{up} = \varepsilon_s \, \sigma \, T_s^4 \qquad (3)$$

Where ε_s is surface broadband emissivity with value of 0.96 (Arya, 2001) and T_s is surface temperature, σ is the Stefan–Boltzmann constant (5.67×10^{-8} W m^{-2} K^{-4}).

$$LW_{net} = \frac{1-\varepsilon_{a,clr}}{\sigma} LW_{up} \, c \qquad (4)$$

Where $\varepsilon_{a,clr}$ is effective clear sky atmospheric emissivity, and *c* represents cloud-free sky conditions. Generally, atmospheric effective emissivity $\varepsilon_{a,clr}$ under clear sky conditions is based on the ambient temperature and water vapor pressure (VP). Downward longwave radiation (LW_{down}) reaching the surface will increase with increase in cloud cover hence effect of cloud cover was also included in Equation 4. Downward long-wave radiation is critical for accurately estimating surface temperature and surface emissivity (Takur et al., 2022). In a synoptic experiment conducted in Phoenix, Arizona, Idso (1981) demonstrated that variable atmospheric concentrations of water dimers cause water vapor-associated thermal emittance variation in cloudless skies. Therefore, a physical model was developed for the atmosphere's effective emittance (Idso, 1981).

$\varepsilon_{a,clr}$ is calculated from Idso (1981) when VP > 0.5 kpa:

$$\varepsilon_{a,clr} = a_1 + a_2 \left[e_a^{\left(\frac{1500}{T_a}\right)}\right] \qquad (5)$$

Where a_1 is 0.7, a_2 is 5.95 × 10 $^{-5}$ hpa^{-1}, and e_a is average VP. Idso (1981) proposed that air emissivity should be nonlinearly dependent on e_a and T_a and proposed aseparate relation for days when VP is <0.5 kpa. Since > 98% of our data has VP >0.5 kpa, we focused on only one part of equation (5) for WISP evaluation.

2.2.3 Inputs to the WISP Model

The inputs of the WISP model (Diak et al., 1998) include daily average air temperature, vapor pressure, and insolation data. The gridded weather data was obtained from the National Oceanic and Atmospheric Administration (NOAA) NCEP (https://www.nco.ncep.noaa.gov/pmb/products/rtma/). From meteorological data, relative humidity was computed. Model input of daily average temperature and vapor pressure for WISP was based on rectilinear grid interpolation for latitude and longitude. In addition, the UW Space Science insolation model was used to estimate hourly insolation, which estimates solar energy at the surface by comparing each pixel of twelve GOES satellite images with and without clouds. Finally, all observations and model inputs were converted into a daily time scale for comparison. The canopy cover was used to convert WISP PET to ET_a for comparison with observed ET_a. Canopy cover was obtained from field data, and missing data were interpolated and adjusted by comparing it with terra Moderate Resolution Imaging Spectroradiometer (MODIS) GPP product (MYD17A2H), MODIS leaf area index (LAI), and the fraction of photosynthetically active radiation (FPAR) product (MCD15A2H) with 500m, eight-day resolution.

2.2.4 Meteorology and Eddy covariance observations

Meteorological data and land-atmosphere fluxes, such as the latent heat flux of evaporation (LE), were obtained from flux towers deployed on nine observation sites in WCS during 2018-2022 (Table 2.1). Out of nine locations, three sites, registered in the Dept of Energy Ameriflux Network Management Project database as US-CS1, US-CS3, US-CS4, were used to evaluate the WISP model and track the sources of uncertainty (Fig. 2.1). The remaining six sites (US-CC1, US-CC2, US-CS5, US-CS6, US-DFC, US-DFk) were used for validation of the WISP model corrections. Sites DFC and DFK are located to the south, outside the boundary of (WCS). Crops such as potatoes, corn, and forage, e.g., pasture and wheatgrass pasture, were cultivated at the beginning of April and harvested by the beginning of September. Therefore, only the growing season (May-August) was included in this study.

Instrumentation for US-CS1, US-CS3, US-CS4, US-CS5, and US-CS6 includes EC portable trailer-based towers with Campbell-Scientific, Inc. IRGASON dual sonic anemometer (accuracy, 1% standard deviation of calibration residuals, offset error < ± 8.0 cm s-1 for u_x, uy, and < ± 4.0 cm s-1 for u_z) and infrared CO_2 and H_2O gas analyzers (accuracy, 2% standard deviation of calibration residuals, gain drift = ± 0.3% of reading/°C, cross-sensitivity = ±0.1 mol H2O/mol CO2) (https://www.campbellsci.com/irgason). These instruments and basic meteorological sensors (temperature, humidity, pressure) were mounted approximately 1-2 m above the canopy to capture mesoscale flux using a 3-meter tripod for the farm. The models used for these sensors are common, quick, easy to install, and technically accurate. For example, Campbell Scientific CR6 dataloggers (accuracy, ± 0.06% of measurement + offset at -40° to +70°C) were installed for US-

CS1, US-CS3, US-CS4, US-CS5, US-CS6 (https://www.campbellsci.com/cr6). A net radiometer (from Apogee Instruments (SN-500) was used to measure solar radiation (https://www.apogeeinstruments.com/net-radiometer/). In the four-component net radiometer, blackbody pyranometers and pyrgeometers face upwards and downwards to measure incoming and outgoing shortwave and longwave radiation. Water (liquid and frozen) was kept away from each of the four sensors by a 0.2 W heater. Therefore, errors were minimized when dew, frost, rain, or snow blocked the radiation path. Temperature and relative humidity were measured using the Campbell Scientific CS215 T/RH sensor (https://www.campbellsci.com/cs215-l). IRGASON dual sonic anemometer and infrared CO_2 and H2O gas analyzers were installed at the US-DFK site, while Licor 7500, 7700, and a Gill sonic anemometer were installed at the US-DFC site. CC1 and CC2 used LI-7200 IRGASON and CSAT-3 sonic anemometers. The same calibration technique was used for instruments at all sites. The temperature, relative humidity, and solar radiation were measured every 20 seconds and averaged over a 1-minute interval.

Flux measurements were recorded at 10 Hz. Data for validation sites was obtained from the Ameriflux site (Table 2.1). LE was used to compute ET_a data. All observations were averaged on a daily timestep for comparison with WISP model inputs and outputs. Energy balance closure correction was not applied to the fluxes to reduce the sources of uncertainty. The post-processing software REddyProc (Wutzler et al. 2018) was used to fill gaps in quality-controlled half-hourly ET observations. REddyProc method uses temporal autocorrelation and co-variation of turbulent

fluxes, and gaps were filled based on available information about incoming solar radiation, air temperature, and vapor pressure deficient based on marginal distribution sampling.

2.2.5 Approaches to evaluation and correction of WISP model

WISP model equations were coded and run in python 3.9 to evaluate uncertainty in model inputs and empirical assumptions. Observations of air temperature (^{o}C), vapor pressure (kpa), and insolation (Wm^{-2}) were compared with WISP model inputs to assess the uncertainty in model inputs.

Table 2.1: General Site Information and Characteristics for training and validation sites

Site name	Duration	Crop type	Latitude (N), Longitude (E)	Elevation	Training/Validation Status	Sample size (n)	Doi
US-CS1	2018	Potatoes	44.10, -89.54	328	Training	63	Desai (2021), https://doi.org/10.17190/AMF/1617710
US-CS3	2019	Potatoes	44.14, -89.57	328	Training	123	Desai (2021), https://doi.org/10.17190/AMF/1617713
US-CS4	2020	Potatoes	44.16, -89.55	328	Training	123	Desai (2021), https://doi.org/10.17190/AMF/1756417
US-CC1	2021	Corn	44.07, -89.68	314	Validation	38	Stoy et al. (2022), https://doi.org/10.17190/AMF/1865475
US-CC2	2021	Corn	44.11, -89.62	320	Validation	18	Stoy et al. (2022), https://doi.org/10.17190/AMF/1865476
US-CS5	2021	Potatoes	44.11, -89.54	328	Validation	105	Desai (2022), https://doi.org/10.17190/AMF/1846663
US-CS6	2022	Potatoes	44.43, -89.54	328	Validation	64	Desai (2022), https://doi.org/10.17190/AMF/1846663
US-DFC	2019	Pasture	43.34, -89.71	256.9	Validation	68	Duff & Desai (2022), https://doi.org/10.17190/AMF/1660340
US-DFK	2019-2020	Wheatgrass	43.34, -89.72	210	Validation	184	Duff & Desai (2022), https://doi.org/10.17190/AMF/1825937

WISP LW_{net} radiation (equation 4) were also compared with measurements to evaluate model assumptions. For model assumptions, the empirical model (Idso, 1981) for the estimate of effective atmospheric emissivity ($\varepsilon_{a,clr}$) was evaluated by comparing the WISP idso $\varepsilon_{a,clr}$ (equation 5) with measured $\varepsilon_{a,clr}$. Based on the WISP model, during clear sky conditions, upward

LW (LW_{up}) is equal to downward (LW_{down}), so LW_{net} is zero under those conditions. Using that assumption, measured $\varepsilon_{a,clr}$ was computed using observed LW_{up} and LW_{down} (equation 6) for clear sky conditions.

$$Measured\ \varepsilon_{a,clr} = \varepsilon_s \frac{LW_{down}}{LW_{up}} \quad (6)$$

Clear sky conditions were determined using Hulstrom et al., (1985) model for clear sky irradiance. That model is based on empirical representations of radiative transfer equations and attempts to take into account the cumulative effects of water vapor, ozone, aerosols and other gases, and Rayleigh (molecular) scattering upon sunlight reaching Earth's surface. The observed SW_{in} during clear sky conditions was compared with clear sky irradiance at Earth's surface to compute days with clear sky conditions. Then measured $\varepsilon_{a,clr}$ was calculated only for days with clear skies. Ordinary least square method (equation 7) was used to calculate intercept (b) or a_1 and slope (m) or a_2 by setting up $y = Measured\ \varepsilon_{a,clr}$ (equation 9). $Measured\ \varepsilon_{a,clr}$ was computed using observations from three sites called training sites (US-CS1, US-CS3, US-CS4).

$$y = mx + b \quad (7)$$

$$x = [e_a^{(\frac{1500}{T_a})}] \quad (8)$$

$$Measured\ \varepsilon_{a,clr} = a_1 + a_2 [e_a^{(\frac{1500}{T_a})}] \quad (9)$$

In the WISP model, equation 5 was corrected based on updated a_1 and a_2 values. Then LW_{net} and ET_a from corrected WISP model were compared with observations.

2.2.6 Model evaluations

There are nine sites with three training sites and six validation sites, with a total of five years (growing season May-September) of observations, or 869 LW_{net} daily observations and 786 ET_a daily observations. Three of the nine sites were used for training such that the observations (n=309) from those sites were used for correction of $\varepsilon_{a,clr}$. The same corrected coefficients for $\varepsilon_{a,clr}$ were applied on the validation sites (n=560). For statistical analysis, the coefficient of determination (R^2), Pearson correlation coefficient, Nash–Sutcliffe efficiency (NSE), Willmott's skill score or index of model performance (Willmott, 1981), mean absolute error (MAE), root mean square error (RMSE), and percentage bias (pbias) were used to assess the performance of WISP before and after correction.

2.3 Results

2.3.1 ET and meteorological observations comparison to WISP

Observations from three sites (US-CS1, US-CS3, US-CS4) were compared with WISP model inputs (air temperature, VP, solar insolation) in Fig. 2.2. WISP air temperature and VP agreed well with the measured air temperature and VP, respectively, with Willmott's index of agreement close to 1. The MAE was 0.66 $^oC\ d^{-1}$ for air temperature and 0.08 kpa d^{-1} for VP (Table 2.2). High agreement between SW_{in} and WISP insolation is visible in daily time series (Fig. 2.2b) with NSE of 0.90 and pbias of less than 4%.

2.3.2 LW_{net} comparison before and after correction of WISP model

The mean observed LW_{net} was -64.8 W m^{-2} d^{-1} (range -129 to -0.4 W m^{-2} d^{-1}) for the 2018-2020 growing season (May-Aug), while WISP predicted mean LW_{net} was -24.2 W m^{-2} d^{-1} (-67.3 to -0.6 W m^{-2} d^{-1}). Therefore, the WISP model underestimated LW_{net} with pbias of -62.6% and RMSE of 43.5 W m^{-2} d^{-1} (Fig. 2.3a, Table 2.2). This lack of agreement between observations and model was tracked back to effective clear sky atmospheric emissivity ($\varepsilon_{a,clr}$) (Fig. 2.3b).

The mean observed LW_{net} was -64.8 W m^{-2} d^{-1} (range -129 to -0.4 W m^{-2} d^{-1}) for the 2018-2020 growing season (May-Aug), while WISP predicted mean LW_{net} was -24.2 W m^{-2} d^{-1} (-67.3 to -0.6 W m^{-2} d^{-1}). Therefore, the WISP model underestimated LW_{net} with pbias of -62.8% and RMSE of 43.5 W m^{-2} d^{-1} (Fig. 2.3a, Table 2.2). This lack of agreement between observations and model was tracked back to effective clear sky atmospheric emissivity ($\varepsilon_{a,clr}$) (Fig. 2.3b).

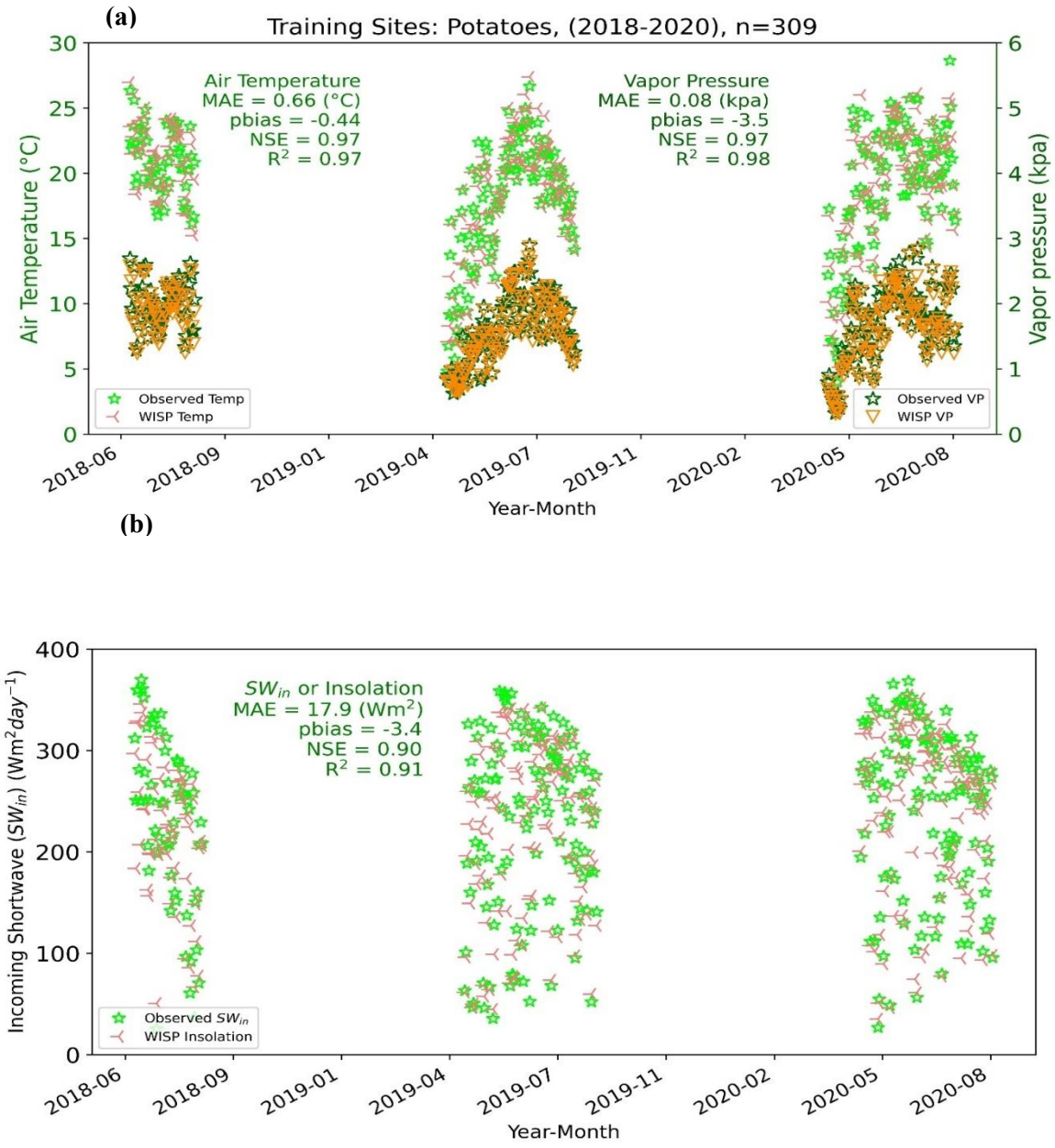

Figure 2.2: (a) Comparison of observed air temperature (°C) and VP (kpa) with WISP air temperature and VP respectively. A secondary y-axis is added for VP; (b) Observed SW_{in} (Wm^{-2}) is also compared with WISP insolation.

The new coefficient value for a_1 is 0.544 and a_2 is 6.4 * 10 $^{-5}$ hpa^{-1}. Corrected $\varepsilon_{a,clr}$ is given in equation 10.

$$\varepsilon_{a,clr} = 0.544 + 6.4 \times 10^{-5} [e_a^{(\frac{1500}{T_a})}] \quad (10)$$

After correcting $\varepsilon_{a,clr}$, LW_{net} was compared with observed LW_{net} for three training sites (Fig. 2.4a). The $\varepsilon_{a,clr}$ correction improved the agreement between WISP and observations for LW_{net} by reducing the pbias from -62.8 to -6.2% (Fig. 2.4b). RMSE was also reduced from 43.5 W m^{-2} d^{-1} to 10.4 W m^{-2} d^{-1}. WISP corrected mean LW_{net} was -60.7 W m^{-2} d^{-1} (range -9.6 to -116 W m^{-2} d^{-1}). LW_{net} improvement was also evident for validation sites (Fig. 2.4c) for all evaluation metrics. For example, for the growing season (2019-2022), the mean observed LW_{net} was -56.1 W m^{-2} d^{-1} (range -122.1 to -3.3 Wm^{-2} d^{-1}). Before the correction, WISP predicted a mean LW_{net} of -23.6 W m^{-2} d^{-1} (range -65.7 to -1.2 W m^{-2} d^{-1}). However, WISP predicted mean LW_{net} was improved after correction with a mean value of -60.1 W m^{-2} d^{-1} (range -116.2 to -8.2 W m^{-2} d^{-1}). Therefore, WISP LW_{net} pbias was reduced from -56.1 to 14.7 (Fig. 2.4 c, d), and RMSE was decreased from 33.1 to 15.4 W m^{-2} d^{-1} after correction (Table 2.3), which improved model precision and accuracy.

Table 2.2: Evaluation statistics are shown for comparison of observations and WISP inputs (n=309)

Evaluation statistics	R^2	Pearson Corr. (r)	NSE	Willmott's index of agreement	MAE	RMSE	pbias
Air Temperature (^oC)	0.97	0.99	0.97	0.99	0.66	0.85	-0.44
Vapor pressure (VP in kpa)	0.98	0.99	0.97	0.99	0.08	0.099	-3.5
Incoming shortwave (SW_{in} in Wm^{-2})	0.91	0.96	0.90	0.97	17.9	26.6	-3.4
Net Long-wave (LW_{net} in Wm^{-2})	0.78	0.88	-1.60	0.56	40.6	43.5	-62.6

2.3.3 WISP model evaluation for ET_a comparison before and after correction

Observed ET_a for training sites ranged from 0.5-5.7 mm d^{-1} (mean 2.9 mm d^{-1}). Predicted ET_a from WISP ranged from 0.54 to 7.28 mm d^{-1} (mean ET_a 4.37 mm d^{-1}). Therefore, before correction, WISP overestimated ET_a with a MAE of 1.15 mm d^{-1} pbias of 19.6 % and NSE of -0.08 (Fig. 2.5a). However, after correction, the obvious agreement can be seen between WISP ET_a and observed ET_a with reduced MAE of 0.72 mm d^{-1}, pbias of -10.1and NSE of 0.52. WISP-corrected ET_a ranged from 0.4-5.6 mm d^{-1} (mean ET_a 3.3 mm d^{-1}). The timeseries plot shows that WISP correction helped reduce the overestimated ET_a for both training and validation sites (Fig. 2.5). Willmott's index of agreement also improved from 0.65 to 0.72, and pbias was reduced from 35% to 1.34 (Table 2.3). The improvement in evaluation statistics was consistent for 2018-2020.

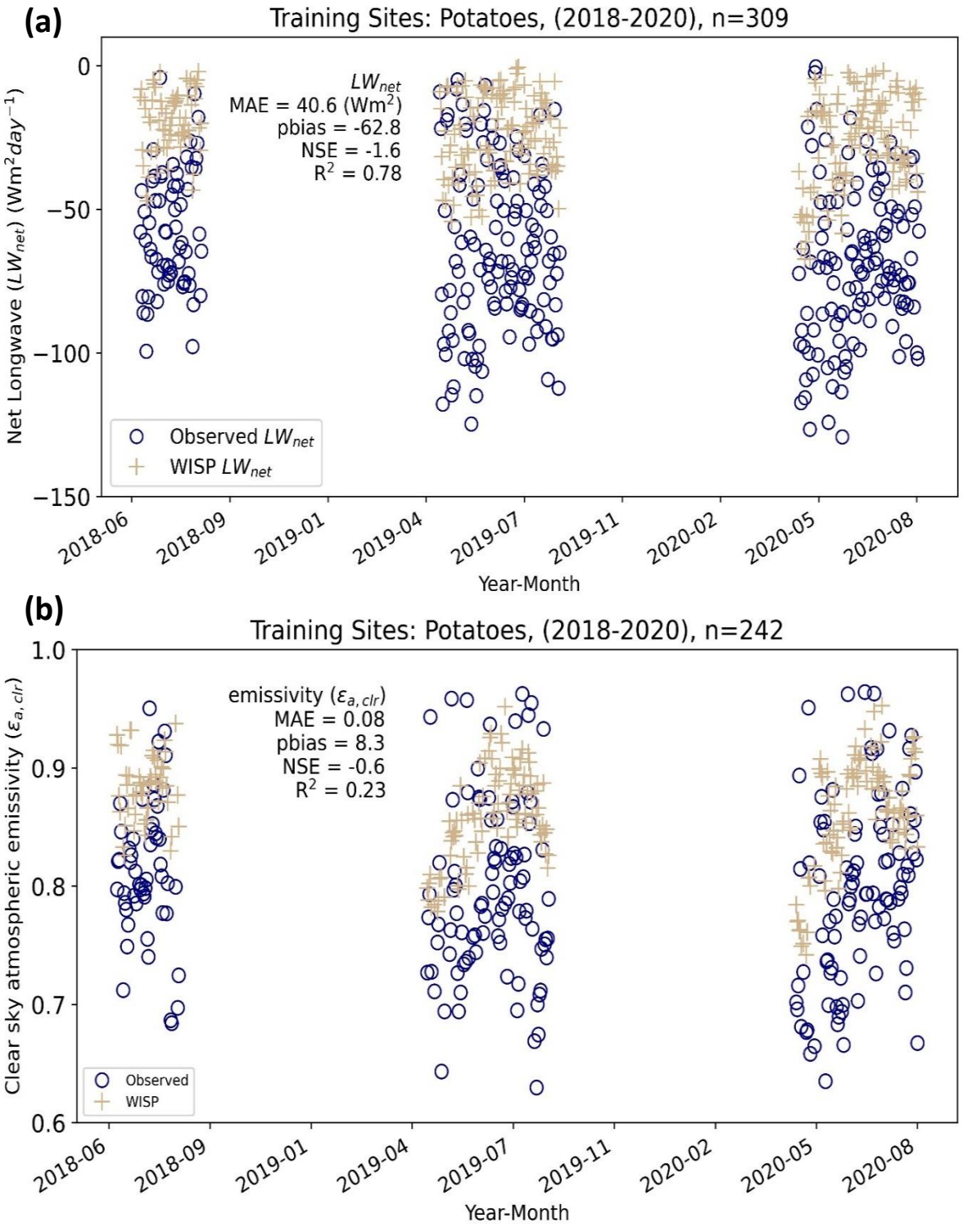

Figure 2.3: LW_{net} and clear sky atmospheric emissivity ($\varepsilon_{a,clr}$) comparison for training sites (a) WISP LW_{net} radiation is compared with observations; (b) $\varepsilon_{a,clr}$ is compared for WISP model and EC observed effective clear sky atmospheric emissivity

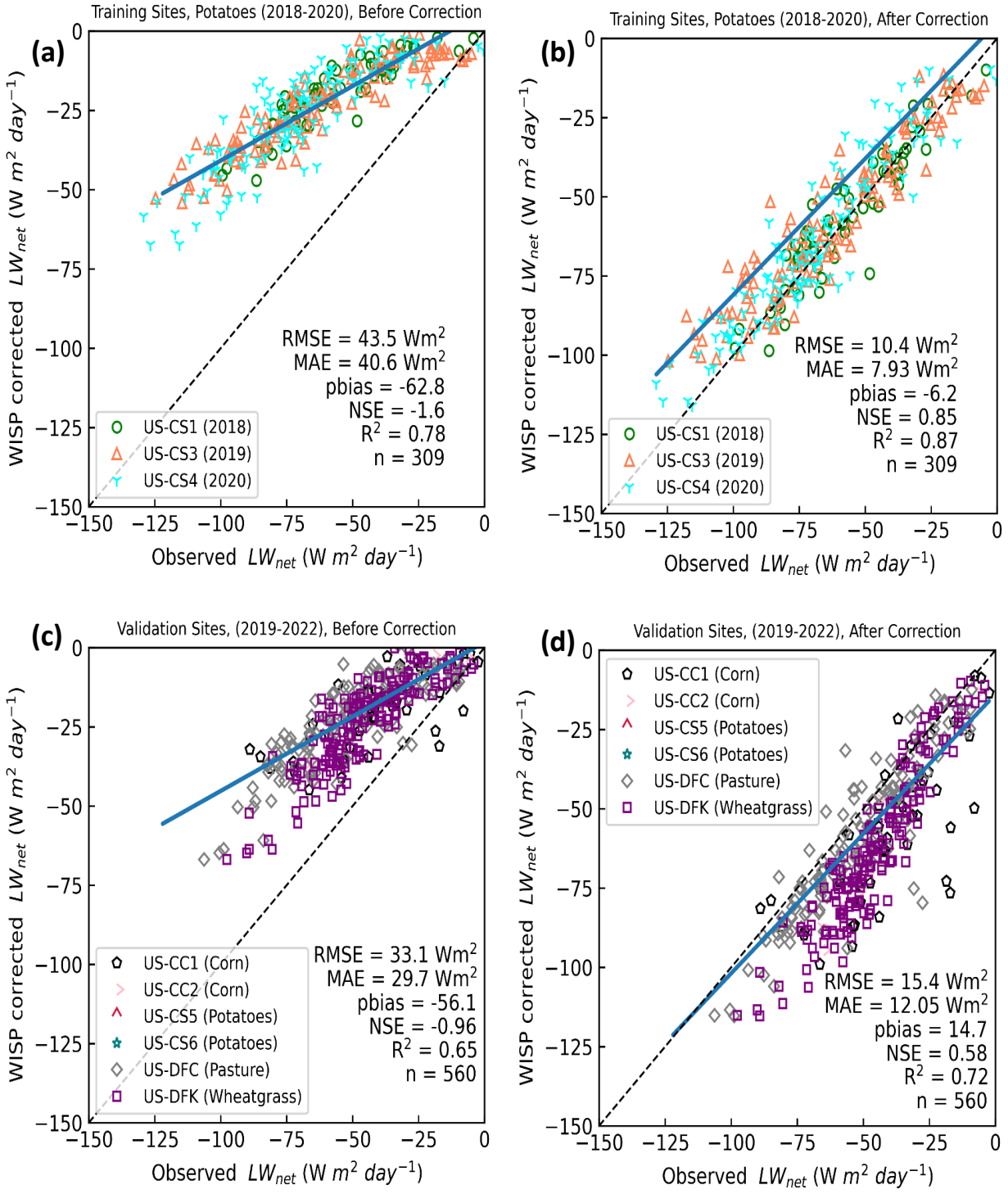

Figure 2.11: (a) WISP LW_{net} was compared with observed LW_{net} before correction for training sites b) WISP corrected LW_{net} compared with observed LW_{net} c) LW_{net} comparison for validation sites is shown d) Corrected LW_{net} is compared with observation for validations sites. Different colors and markers were used to represent different sites. 1:1 line (dotted) and fitted lines are also presented. LW_{net} is compared with observation for validations sites. Different colors and markers were used to represent different sites. 1:1 line (dotted) and fitted lines are also presented.

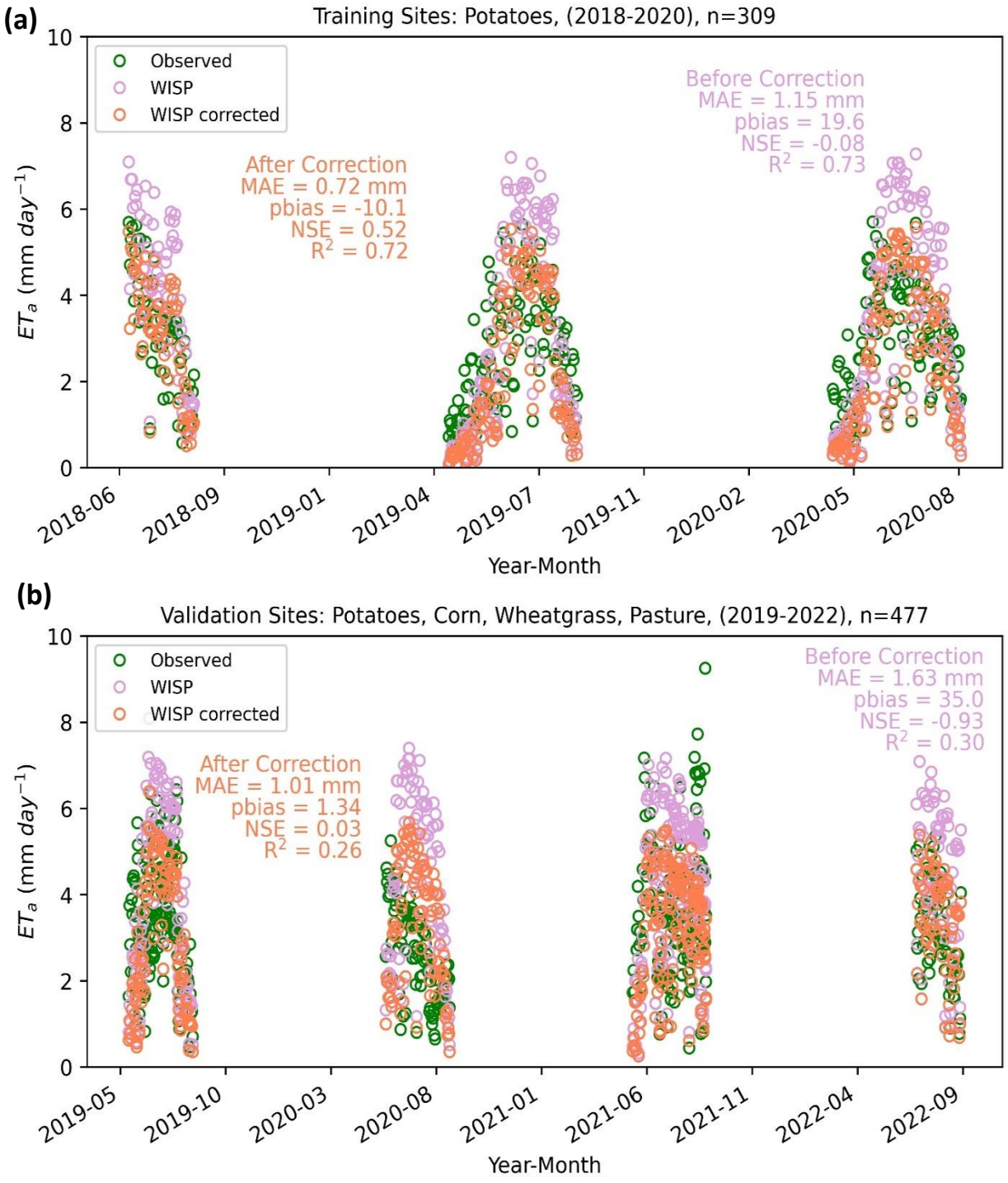

Figure 2.5: (a) ET_a comparison between WISP and observations is shown before and after correction for training sites. b) WISP ET_a before and after correction is compared with observations from validation sites. Evaluation metrics are also shown for comparison.

Table 2.3: Evaluation statistics are shown for comparison of observations and WISP outputs such as LW_{net} and ET_a. for training and validation sites. Evaluation matrices are shown before and after correction.

Observed LW_{net} vs WISP corrected LW_{net} (W m^{-2} d^{-1})	Training (n=309)	0.87	0.94	0.85	0.96	7.98	10.4	-6.45
Observed LW_{net} vs WISP LW_{net} (Wm^{-2} d^{-1})	Validation (n=560)	0.65	0.80	0.96	0.60	29.7	33.1	-56.1
Observed LW_{net} vs WISP corrected LW_{net} (W m^{-2} d^{-1})	Validation (n=560)	0.72	0.85	0.58	0.90	12.05	15.4	14.7
Observed ET_a vs WISP ET_a (mm d^{-1})	Training (n=309)	0.73	0.85	-0.08	0.84	1.15	1.37	19.6
Observed ET_a vs WISP corrected ET_a (mm d^{-1})	Training (n=309)	0.72	0.85	0.52	0.90	0.72	0.91	-10.1
Observed ET_a vs WISP ET_a (mm d^{-1})	Validation (n=477)	0.30	0.54	-0.93	0.65	1.63	1.95	35.0
Observed ET_a vs WISP corrected ET_a (mm d^{-1})	Validation (n=477)	0.26	0.51	0.03	0.72	1.01	1.40	1.34

2.3.3 WISP model evaluation for ET_a comparison before and after correction

Observed ET_a for training sites ranged from 0.5-5.7 mm d^{-1} (mean 2.9 mm d^{-1}). Predicted ET_a from WISP ranged from 0.54 to 7.28 mm d^{-1} (mean ET_a 4.37 mm d^{-1}). Therefore, before correction, WISP overestimated ET_a with a MAE of 1.15 mm d^{-1} and NSE of -0.08 (Fig. 2.5a). However, after correction, obvious agreement can be seen between WISP ET_a and observed ET_a with reduced MAE of 0.72 mm d^{-1} and NSE of 0.52. WISP-corrected ET_a ranged from 0.4-5.6 mm d^{-1} (mean ET_a 3.3 mm d^{-1}). The scatter plot shows that WISP correction helped reduce the overestimated ET_a for both training and validation sites (Fig. 2.5). Willmott's index of agreement also improved from 0.65 to 0.72 and pbias was reduced from 35% to 1.34 (Table 2.3). The improvement in evaluation statistics was consistent for 2018-2020.

Fig. 2.6 shows the variation in ET_a for training sites each year during the growing season. WISP ET_a was consistently higher than observations, especially during June-July. However, after correction, an increase in overlap of WISP-corrected ET_a with observations demonstrates improvement in model performance. For example, WISP overestimated ET for the US-CC1 corn site during June and July. The observed mean ET_a for the US-CS1 site was 3.1 mm d^{-1} (range 0.6-5.6) for 2018 (July-Aug) while WISP-estimated ET_a was 4.1 mm d^{-1} (range 0.7-7.0). After correction, WISP estimated ET_a to be 3.1 mm d^{-1} (range 0.5-5.4) (Table 2.4). After correction, NSE improved remarkably from 0.06 to 0.81 for the US-CS1 potatoe site (Table 2.4). Additionally, MAE was reduced from 1.12 mm d^{-1} to 0.49 mm d^{-1}, and pbias decreased from 33.3% to -0.5%. For the US-CS3 site, although correlation did not improve between WISP and observed ET_a after correction, RMSE was reduced from 1.35 mm d-1 to 0.99 mm d^{-1}. The mean observed ET_a was 2.9 mm d^{-1} (0.5-5.7) for US-CS4 in the growing season of 2020. Even though WISP predicted higher ET_a with a mean of 3.4 mm d^{-1} (0.2-7.3), the agreement between observations and WISP improved with mean ET_a of 2.6 mm d^{-1} (0.1-5.6) after correction. Improvement in WISP performance was also evident with reduced MAE and RMSE (Table 2.4).

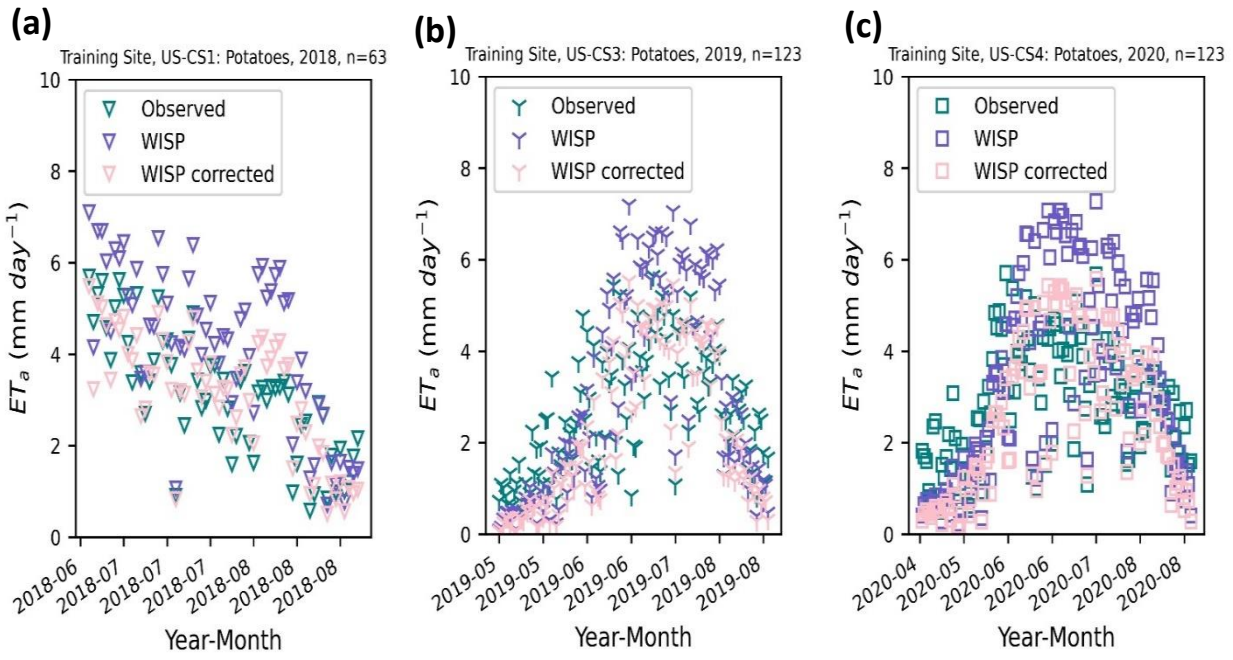

Figure 2.6: a) Observed and WISP ET_a comparison for potatoes in 2018 growing season (July-Aug) b) ET_a comparison for US-CS3 site in 2019 growing season (May-Aug). c) WISP ET_a before and after correction is compared with observations for US-CS4 site for 2020 growing season (May-Aug). The variety of potatoes was same for 2018 and 2020. A different variety of potatoes was planted in 2019.

Fig. 2.7 shows ET_a variability for the growing season for different crops at validation sites in WCS. Statistical analysis of the observed versus WISP ET_a before and after correction is demonstrated in Table 2.5. The mean observed ET_a for corn is reported to be 3.3 mm d^{-1} (range 1.2-7.2) for site US-CC1 in 2021. WISP overestimated ET_a for corn, especially during June and July, with a mean of 5.0 mm d^{-1} (range 2.2-6.8). After correction, mean ET_a was reduced to 3.7 mm d^{-1} (range 0.8-5.4), and pbias was improved from 49.6% to 12.4%, even though the correlation was reduced. After data preprocessing for US-CC2, there were only eighteen days of observations (mid-end of Aug) for model evaluation, and contrary to the expectation, WISP underestimated corn ET_a, and model performance became worse after correction (Fig. 2.7b). On the other hand,

as expected, WISP overestimated potato ET_a at US-CS5 (Fig 4c). The mean observed ET_a at that site was 3.4 mm d^{-1} (range 0.4-7.2). WISP-estimated ET_a was 4.2 mm d^{-1} (range 0.3-7.2) with pbias of 25.1%. After correction, WISP mean ET_a became closer to the observation (3.2 mm d^{-1}, range 0.3-5.5) with reduced pbias of -6.4%. MAE was also reduced from 1.34 mm d^{-1} to 0.75 mm d^{-1}. US-CS6 is another validation site with potatoes (Fig. 2.7d, where WISP overestimated ET_a with a mean of 4.6 mm d^{-1} (range 1.0-7.1) with Willmott's index of agreement of 0.7. After correction, WISP ET_a (mean=3.4 mm d^{-1}, range 0.7-5.4) was in greater agreement with observation (mean=3.1 mm d^{-1}, range 0.8-5.4) with Willmott's index of agreement of 0.89 and a slight improvement in correlation (Table 2.5).

Table 2.4: Correlation metrics (R^2, Pearson correlation), NSE, Willmott's accuracy index, error metrics (MAE, RMSE), and pbias was computed for evaluation of WISP ET_a before and after correction. Sample size and growing season year is also presented. Only the growing season was included. All training sites measured potatoes.

Site	R^2	NSE	Willmott	Pearson Corr. (r)	MAE (mm d^{-1})	RMSE (mm d^{-1})	pbias	n	year
US-CS1 (Without correction)	0.80	0.06	0.83	0.89	1.12	1.3	33.3	63	2018
US-CS1 (With correction)	0.81	0.81	0.95	0.90	0.49	0.59	-0.5	63	2018
US-CS3 (Without correction)	0.74	-0.03	0.85	0.86	1.1	1.35	12.8	123	2019
US-CS3 (With correction)	0.73	0.45	0.89	0.85	0.82	0.99	-15	123	2019
US-CS4 (Without correction)	0.70	-0.26	0.82	0.84	1.2	1.41	18.3	123	2020
US-CS4 (With correction)	0.70	0.4	0.88	0.84	0.75	0.97	-11.2	123	2020

Aside from potatoes and corn, one validation site had pasture measurements. For example, for US-DFC (Fig. 2.7e), improvement was seen in NSE from -0.0005 to 0.24, with a reduction in RMSE from 1.25 to 1.09 mm d^{-1}. The US-DFK site was planted with perennial intermediate wheatgrass (Kernza™, *Thinopyrum intermedium*) in 2019-2020. Compared to observations (mean=2.8 mm d^{-1}, range 0.5-4.8), WISP overestimated intermediate wheatgrass ET_a (mean=4.1 mm d^{-1}, range 0.5-8.1), especially during July and August. However, improvement was demonstrated after correction with a mean observed ET_a of 3.1 mm d^{-1}, (range 0.4-6.4). That also helped to reduce pbias from 48.4% to 12.3% and RMSE from 2.2 to 1.39 mm d^{-1}. For 2020, US-DFK observed mean ET_a was 2.6 mm d^{-1} (range 0.7-5.3) (Fig. 2.7g). WISP predicted ET_a was 4.4 mm d^{-1} (range 0.5-7.4).

Fig. 2.8 shows box plots for overall statistical matrices for the performance of WISP ET_a for training and validation sites. While correlation did not improve, a significant difference was found between WISP model performance for predicting ET_a before and after correction in terms of NSE ($p=0.017$), MAE ($p=0.012$), RMSE ($p=0.02$) and pbias ($p=0.03$). Similar to training sites, the correlation between observations and WISP ET_a was similar before and after correction (Fig. 2.8b). However, a slight decrease can be seen in MAE and RMSE. Overall, pbias was significantly reduced ($p=0.035$) for validation sites after WISP correction.

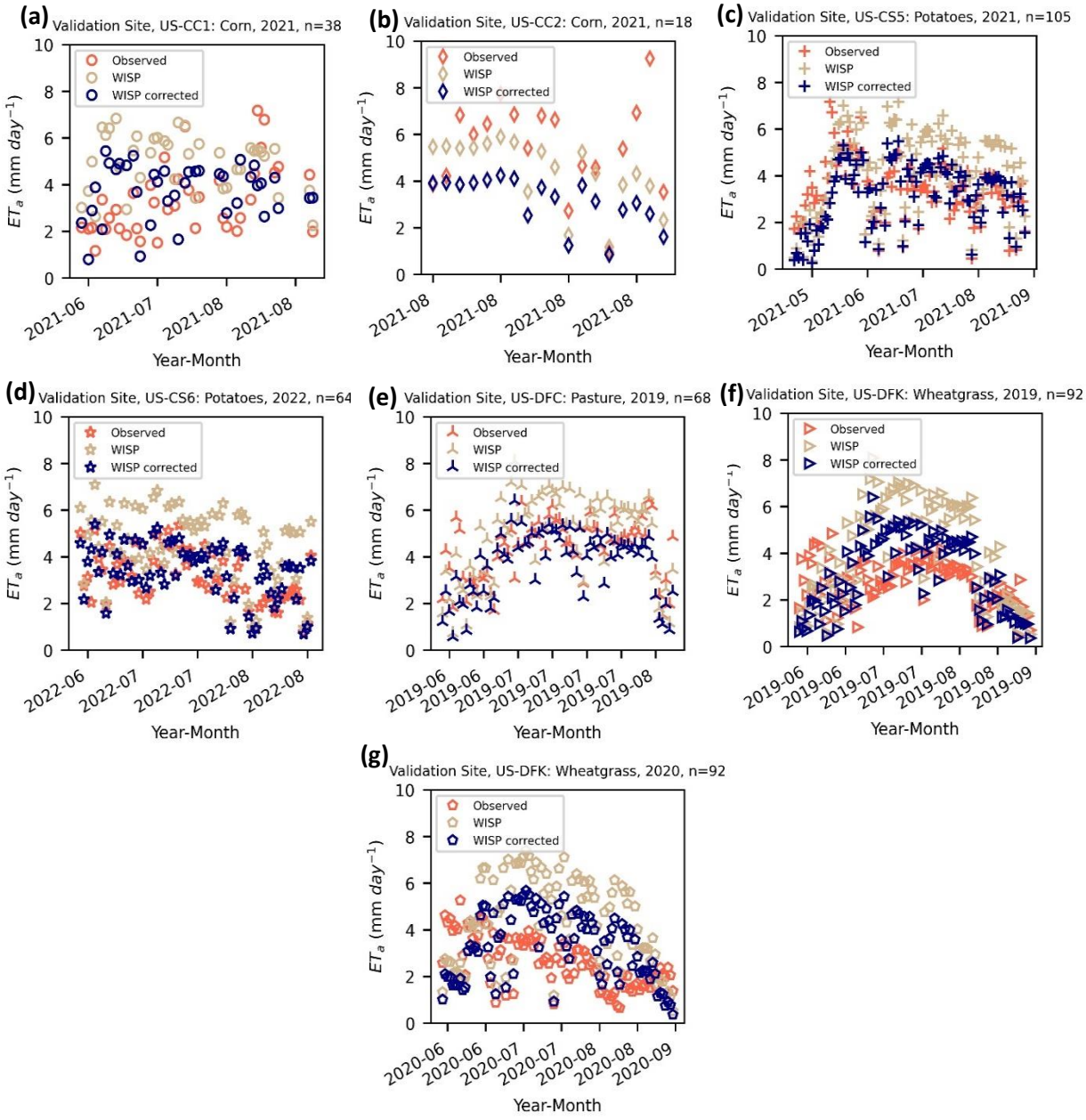

Figure 2.7: WISP ET_a is compared with observed ET_a before and after correction during the growing season (July-Aug) for validation sites a) US-CC1 site was planted with corn in 2021 (June-Aug). b) US-CC2 site was planted with corn in 2021 (data were available for 18 days in Aug). c) US-CS5 site was planted with potatoes in 2021 (May-Aug). d) US-CS6 site was planted with same potatoes variety in 2022 (May-Aug). e) US-DFC site was planted with pasture in 2019 (June-Aug). f) US-DFK site was planted with wheatgrass in 2019 (June-Aug). g) US-DFK site was planted with intermediate wheatgrass in 2020 (June-Aug).

Table 2.5: Statistical analysis of the observed versus WISP ET_a before and after correction for the validation sites

Site	R	NSE	Willmott	Pearson Corr. (r)	MAE (mm d^{-1})	RMSE (mm d^{-1})	pbias	n	year	Crop type
US-CC1 (Without Correction)	0.14	-1.26	0.52	0.38	1.94	2.24	49.6	38	2021	Corn
US-CC1 (With Correction)	0.003	-0.59	0.42	0.05	1.6	1.88	12.4	38	2021	
US-CC2 (Without Correction)	0.38	0.048	0.71	0.62	1.5	1.88	-19.9	18	2021	Corn
US-CC2 (With Correction)	0.35	-1.1	0.58	0.59	2.3	2.8	-42.7	18	2021	
US-CS5 (Without Correction)	0.49	-0.43	0.74	0.7	1.34	1.54	25.1	105	2021	Potatoes
US-CS5 (With Correction)	0.48	0.33	0.82	0.69	0.75	1.1	-6.4	105	2021	
US-CS6 (Without Correction)	0.65	-1.25	0.7	0.81	1.5	1.7	47.9	64	2022	Potatoes
US-CS6 (With Correction)	0.68	0.58	0.89	0.82	0.59	0.75	10.2	64	2022	
US-DFC (Without Correction)	0.63	-0.0005	0.83	0.79	1.002	1.25	12.5	68	2019	Pasture
US-DFC (With Correction)	0.6	0.24	0.83	0.78	0.83	1.09	-14.4	68	2019	
US-DFK (Without Correction)	0.32	-4.2	0.52	0.56	1.84	2.2	48.4	92	2019	Wheatgrass
US-DFK (With Correction)	0.31	-1.08	0.68	0.56	1.13	1.39	12.3	92	2019	
US-DFK (Without Correction)	0.14	-4.4	0.44	0.38	2.2	2.5	66.9	92	2019	Wheatgrass
US-DFK (With Correction)	0.16	-1.1	0.6	0.4	1.38	1.56	25.6	92	2020	

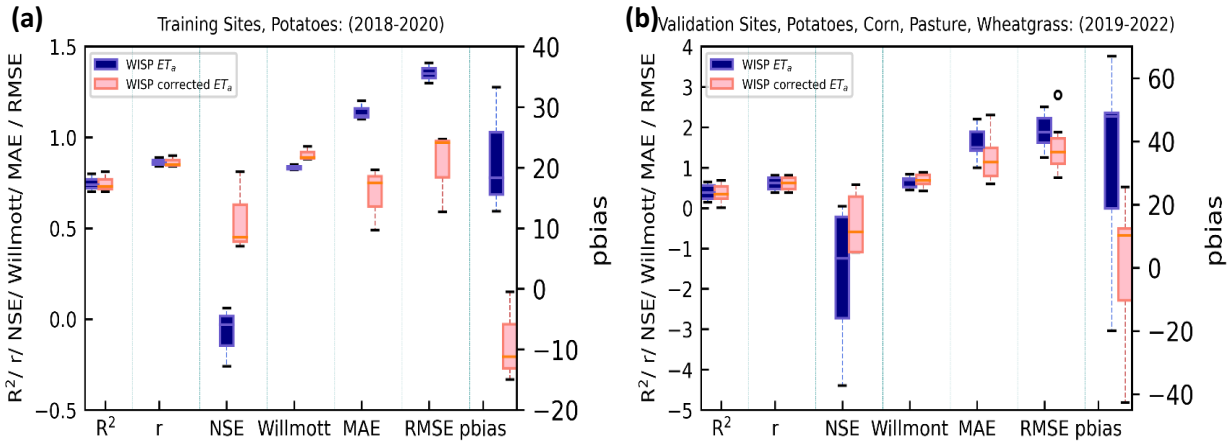

Figure 2.8: Statistical metrics were calculated to evaluate the performance of WISP ET_a before and after correction. Because of difference in scale, pbias was included on the y axis on the right. R^2, r, NSE, Willmott's index of accuracy, MAE and RMSE were included on the y-axis on the left. a) WISP evaluation on training sites, b) WISP evaluation on validations sites

2.3.4 WISP model ET_a evaluation by month

Overall, a higher correlation was found between observed and WISP ET_a during the latter part of the growing season (July-Aug) (Fig. 2.9 a,b) for training sites. While correlation did not differ significantly ($p>0.05$) between WISP ET_a before and after correction, NSE was significantly improved during May, June and August ($p=0.042$, 0.009, 0.046) after correction (Fig. 2.9c). That also helped to reduce MAE and RMSE for July ($p=0.014$, 0.018) and August ($p=0.02$,0.02) (Fig. 2.9 e,f) and improved model precision and accuracy. Before correction, WISP model ET_a had a higher

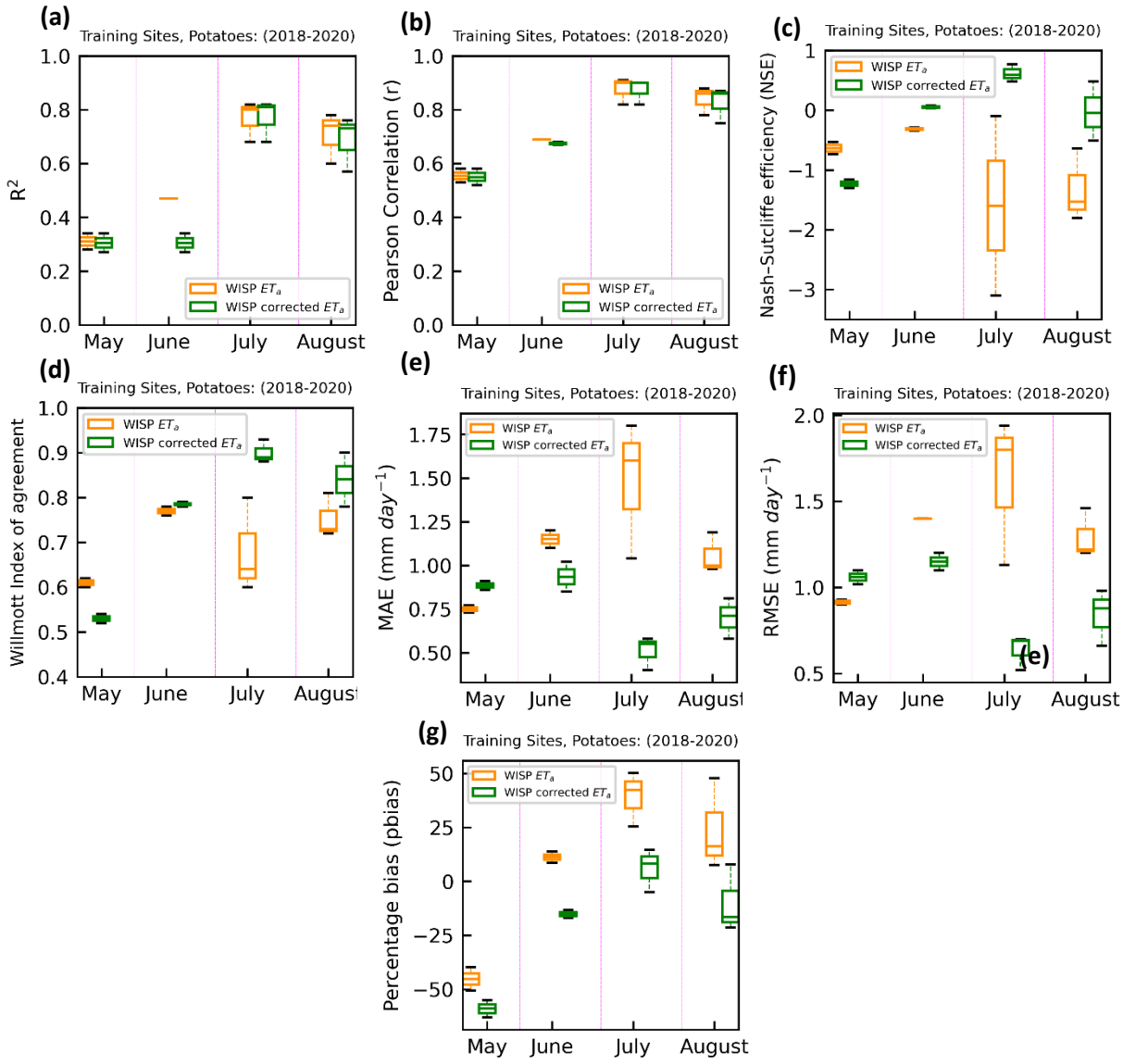

Figure 2.9: Effect of variability in ET_a during the growing season on performance of WISP before and after correction for all training sites. Evaluation statistics such as a) R2, b) Pearson correlation (r), c) NSE d) Willmott's index of agreement, e) MAE, f) RMSE, g) pbias are calculated.

range for NSE, Willmott's index, MAE and RMSE. Correction helped to reduce this variability across sites. Additionally, pbias was significantly reduced (p=0.02) for July.

For validation sites, July and August had a higher range of variability (Fig. 2.10). Additionally, compared to training sites, there was a higher overlap between the performance of

WISP between and after correction. However, by looking at the means of the violin plot, there is an improvement in NSE (Fig. 2.10c) and Willmott's index of agreement (Fig. 2.10d). Decrease in MAE and RMSE can also be seen, especially for July (p=0.024, 0.04). Correction also improved the pbias variability with a significant reduction for bias in July ET_a (p=0.0001).

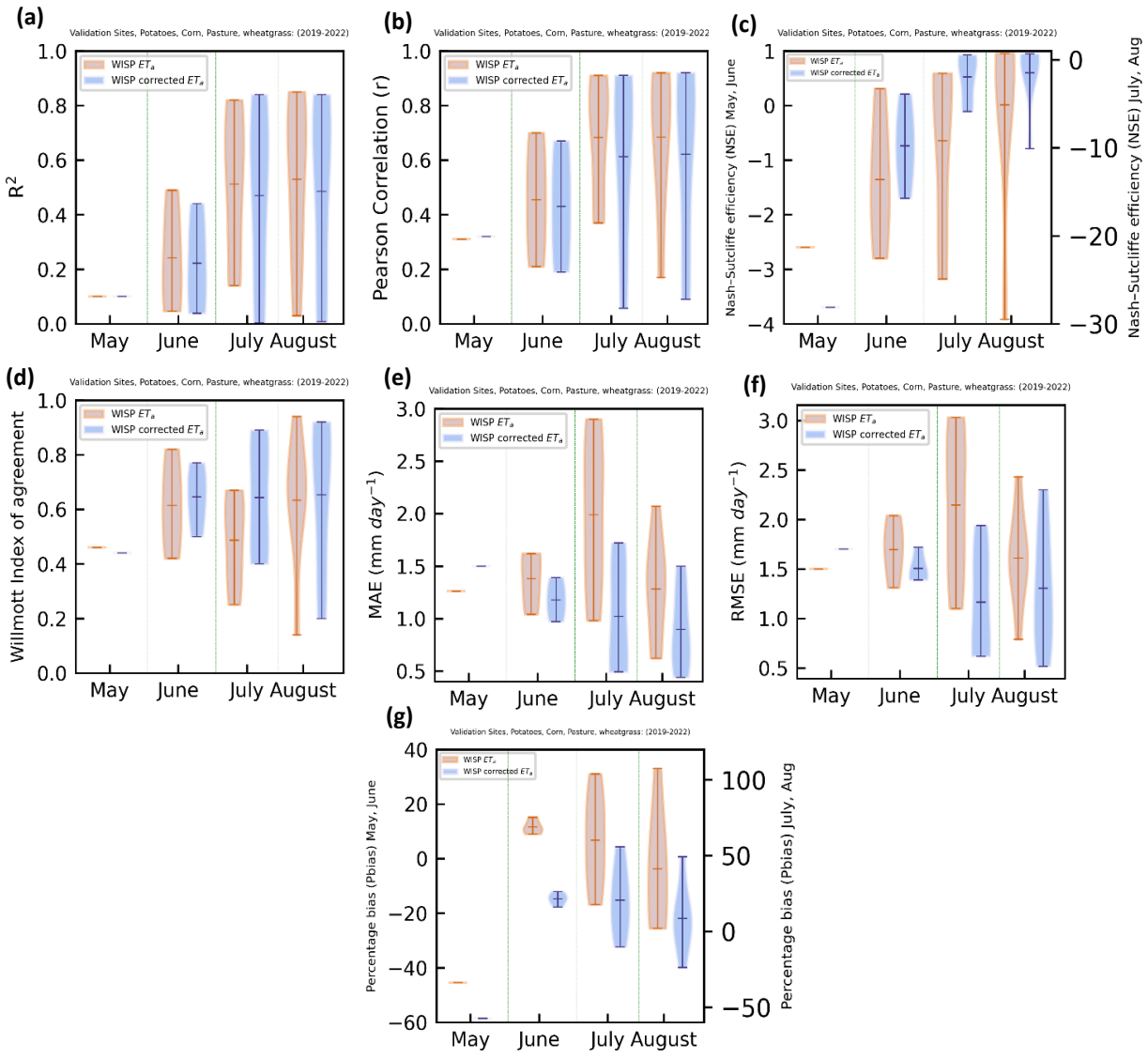

Figure 2.10: Violin plots for overall WISP ET_a performance by month for all validation sites before and after correction. Evaluation statistics are measured for all validation sites a) R2, b) pearson correlation (r), c) NSE d) Willmott's index of agreement, e) MAE, f) RMSE, g) pbias

2.4 Discussion

2.4.1 Role of emissivity

A better understanding of evapotranspiration rates is necessary because it is the largest component of the terrestrial water budget after precipitation. The objective of this study was to evaluate and correct a regional ET model WISP. According to the results, in general, the effective clear sky atmospheric emissivity ($\varepsilon_{a,clr}$) correction for the WISP model successfully reduced uncertainty in LW_{net} and ET_a. With correction, R^2 is the same as the original WISP. Original and corrected WISP had similar performance in capturing day-to-day ET variation, but corrected WISP was better in magnitude, which makes sense in our energy-limited system.

An important finding was that the specified values coefficients in Idso's model (1981) calibrated in a water-limited region were changed from a_1= 0.7 and a_2 =5.95 $\times 10^{-5}$ hpa^{-1} *to* a_1= 0.544 and a_2 =6.4 $\times 10^{-5}$ hpa^{-1}, respectively, in an energy-limited region. Errors associated with LW_{net} and ET_a were much smaller after Idso $\varepsilon_{a,clr}$ correction for LW_{net} in WISP. Since the data used to develop the Idso equation were obtained in a water-limited ecosystem with different atmospheric composition, including dust, the differences between calibrated and original Idso coefficients might be explained by atmospheric turbidity (variations in atmospheric dust content can significantly affect longwave irradiance downward) (Alados-Arboledas et al., 1986). Like the greenhouse effect, airborne dust absorbs and scatters sunlight entering the atmosphere, reducing the amount that reaches the surface and absorbing and re-emitting long-wave radiation that bounces back up from the surface. As a result of shortwave heating, dust causes net heating.

For example, dust in the Saharan Air Layer has a negative top-of-atmosphere direct radiative effect (DRE) (Ryder, 2021). Hence in a water-limited region, dust variability is dominant in driving atmospheric heating and surface radiation (Marsham et al., 2016).

LW_{down} is modeled by the Idso empirical model as a function of surface air temperature (T_a) and vapor pressure (e_a): $LW_{down} = f(T_a, e_a)$. Other models have been compared with Idso's model for estimating LW radiation. For example, the comparison of LW radiation estimation methods based on air temperature (Weiss, 1982, Jensen et al., 1990) and methods based on air and soil surface temperatures (Brutsaert, 1981, Ortega-Farias et al., 2000, Saito et al., 2006, Cui et al., 2010, Hemmati et al., 2012) showed that they underestimate air emissivity during night time and in cloudy conditions (Howell et al., 1993, Choi et al., 2008). Compared with methods that only consider air temperature, Idso's model (1981) showed the highest performance since it considered both vapor pressure and air temperature. Precipitable water and humidity have also been identified as the main indicators of radiatively active water vapor based on other studies (Hatfield et al., 1983; Prata, 1996; Flerchinger et al., 2009; Juszak and Pellicciotti, 2013). Hence, water vapor or humidity is the most significant absorber and emitter of longwave radiation in clear sky conditions.

Longwave radiation and sensible heat flux observations are crucial for estimating surface temperature and emissivity (Thakur et al., 2022). As empirical models (e.g., the Idso model) are based on only a few ground observation sites, they tend to have greater errors in other sites. Therefore, it is necessary to evaluate and compare those models' performance. For example, a study evaluating 13 algorithms, including the Idso model (1981), found that Ångström (1915),

Dilley and O'Brien (1998), and Prata (1996) algorithms excelled in predicting LW_{down} under clear skies (Flerchinger et al., 2008). Wang and Dickinson (2013) found that the atmospheric emissivity calculated by Idso (1981) under hot and humid conditions was larger than unity, which is physically impossible. It has also been found that the Idso model performed well in humid conditions but overestimated LW_{down} in dry conditions (Aguiar et al. 2011; Wu et al., 2012). On the other hand, we found that the Idso model underestimated LW_{net} in our humid study area.

Studies were done to locally calibrate Idso coefficients using data from the application location to improve the accuracy of empirical model estimations (Table 2.6). For example, in one study done over a flat area without shady spots in Valladolid, Spain, the Idso coefficients were updated to a_1= 0.66 and a_2 = 3.92 *× 10 $^{-5}$ hpa^{-1} (Bilbao and Miguel, 2007). In another study by Juszak and Pellicciotti (2013), Idso's model parameterization was done on low and high-elevation glaciers at Haut Glacier d'Arolla. The calibrated values for that location were* a_1= 0.699 and a_2 =1.2 *× 10 $^{-3}$ hpa^{-1}* for the summer season. The RMSE of the LW_{down} radiation for the training site in that region was around 21.5 W m^{-2}. RMSE. The RMSE for LW_{net} at our training sites was 10.4 Wm^{-2}. That study by *Juszak and Pellicciotti (2013)* also found that parameterizations strongly underestimated the observed variability in LW_{down}, because the measured data have a standard deviation of about 45 W m^{-2}, and the standard deviation obtained with the Idso parameterizations (with Kimball et al., (1982) cloud correction) was around 32 W m^{-2}. Formetta et al., (2016) reported RMSE between 15-25 W m^{-2} when the Idso model was adjusted for 24 FLUXNET sites in North America W m^{-2}. Moreover, in a study by Gua et al. (2019), adjusted Idso model coefficients based on data from FLUXNET data from 71 globally distributed sites (including

3 cropland sites) were a_1= 0.684 and a_2 = 4.69 × *10* $^{-5}$ *hpa*$^{-1}$, and the RMSE of LW_{down} radiation was around 21.21 W m^{-2}. Compared to that, the standard deviation for our measured data on validation sites was 26 W m^{-2}, and the corrected WISP model standard deviation was 25.3 Wm^{-2}. It was also reported in the study of Formetta et al., (2016) that Idso (1981) and Brunt (1932) model parametrization provides the best results and lower variability, independently of the latitude and longitude ranges where they are applied. Table 2.7 shows the comparison of adjusted coefficients for Idso (1981) model from different studies.

Based on the comparison of adjusted coefficients from the present study with previous studies, it is evident that our value of a_1 was lower compared to the other studies (Table 2.7). On the other hand, the adjusted a_2 value was within the range compared to other studies except for the forest in Finland and Haut Glacier d'Arolla, where it was two orders of magnitude larger than our study and other studies. However, there was no common pattern of variations among adjusted values from other studies. Alternate models that account for variations in aerosols, upper atmosphere properties, or surface moisture may be needed to unify parameterizations. The growing and rich network of surface longwave and atmospheric properties observations, such as those made within Fluxnet (Formetta et al., 2016) or the National Oceanic and Atmospheric Administration's (NOAA's) Surface Radiation budget Network (SURFRAD) program (Augustine et al., 2000), provides an opportunity to do so at large spatial scales in future work.

The hydrologic models such as HYDRUS (Saito, 2006; Simunek et al., 2009) and land surface models such as two-source energy balance (TSEB, Norman et al., 1995), surface energy

balance system (SEBS, Su et al., 2002) model, and global data assimilation system (GLDAS, Rodell et al., 2004) model estimate ET based on Idso model. Therefore, the results of this study will have significant theoretical implications for evaluating model fidelity in resolving spatial heterogeneity in latent heat fluxes, which is essential to convective processes and hydrologic modeling.

Table 2.6: **Adjusted coefficients for the Idso model (1981) based on different studies. Site description and climate regime information are also presented.**

a_1	a_2 (hpa-1)	Site Location	Site description	Climate regime	Reference
0.70	5.95 * 10^{-5}	Phoenix, Arizona	Roof of the Water Conservation Laboratory at Phoenix	Tropical and Subtropical Desert Climate	Idso (1981)
0.544	6.4 * 10^{-5}	Wisconsin, USA	Crops	Hot-summer humid, continental climate	Present study
0.72	9 * 10^{-3}	Sodankylä observatory, Finland	Swamp, boreal forests	Warm-summer humid continental climate	Niemelä et al., (2001)
0.66	3.92 * 10^{-5}	Valladolid, Spain	Flat area, without shady spots	Hot-summer Mediterranean climate	Bilbao & Miguel, (2007)
0.699	1.2 * 10^{-3}	Low and high-elevation melting glacier	*Haut Glacier d'Arolla*	Ice cap climate	Juszak & Pellicciotti, (2013)
0.685	3.2 * 10^{-5}	Contiguous United States	Different types of landuse	Different types of climates	Li et al., (2017)
0.684	4.69 *10^{-5}	71 globally distributed sites	Different land use type (including crops)	Different types of climates	Guo et al., (2019)
0.63	7.5*10^{-5}	Tibetan Plateau (TP)	Highland	Cold semi-arid climate	Liu et al., (2020)

2.4.2 Corrected model performance by site and month

Our second key finding is that when WISP is calibrated and validated at different sites using the same coefficients, it models the LW_{net} and ET_a with acceptable accuracy but with a wider range of errors than at the training sites. It could be concluded from these results that energy interaction mechanisms and biophysical processes differ across locations, even in similar climates, due to subtle variations in aerosol optical thickness, cloudiness, or average humidity. A difference in the temperature and moisture profiles in the atmosphere could also explain the scatter in

model performance from one site to another, as recognized by Swinbank (1964). Due to this, models based only on these measurements vary at locations other than the calibration sites. It was also noted in the early studies on longwave irradiance that the spatial variability of LW_{net} and ET_a could not be explained solely by meteorological variables such as air temperature, humidity, and the simple cloud cover estimate commonly used (*Juszak and Pellicciotti, 2013).*

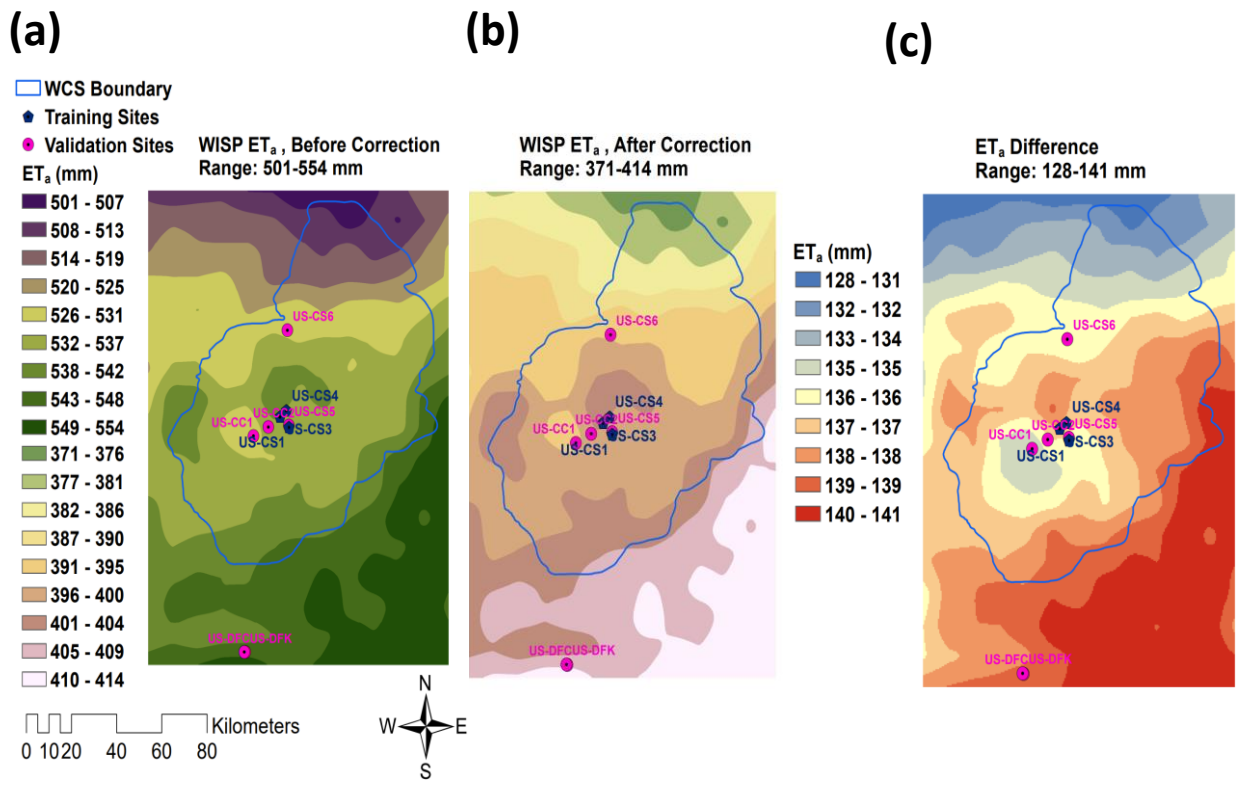

Figure 2.11: Cumulative ET_a for WCS region for the growing season (May-Aug) of 2022. The map was extended beyond the WCS boundary to include DFC and DFK testing sites. (a) WISP cumulative ET_a before correction, (b) WISP cumulative ET_a after correction. ET_a difference map between WISP before and after correction

Fig. 2.11 shows the comparison of WISP ET_a before and after correction. For the growing season (May-Aug) of 2022 in the WCS region, the cumulative WISP ET_a range is 501-544 mm

before correction. However, after correction, the cumulative ET_a was reduced to a range of 371-414 mm. Hence a difference of 128-141 mm for one growing season has an implication for water conservation based on accurate estimates of ET_a. The WISP-optimized approach to ET estimation improved model performance overall, but a few factors can still cause variability in estimates across months. There could also be errors in LW_{net} or ET_a estimation due to heat storage in the water column and wet surfaces on irrigated crops. In the shoulder months, such as early May and late August, we also found large uncertainty in estimating ET from the Priestley-Taylor model. Explicit consideration of soil and plant components of the energy balance is expected to improve estimations of evapotranspiration during that period of the year, even though evapotranspiration during that time of the year contributes to a small portion of total annual water use.

2.4.3 Model limitations and potential improvement

Plant transpiration and soil evaporation are not separated in our approach. This leads to uncertainty in ET estimates during the early stages of growing seasons, when ET is mainly determined by evaporation from the soil. Errors in field measurements of ET can be propagated to the optimization of WISP. Eddy covariance measurements, for example, have been well-recognized for the energy balance closure issue (Twine et al., 2000; Baldocchi et al., 2016; Eichelmann et al., 2018). In the WCS, incomplete storage calculations might cause a lack of energy balance closure at our sites. Additionally, the scale mismatch between WISP ET and tower measurements can introduce additional uncertainties due to the varying footprint size of flux towers. A rigorous footprint approximation is also needed for a fair comparison with field measurements in future studies (Anderson et al., 2018; Wong et al., 2021). In the future, to

determine the weight and extent of the pixel window at flux tower sites, flux footprint model (Kljun et al. 2015) can be implemented. Energy and ET estimates can also be uncertain when water use differs across a region, especially during the rainy season or during plant growth and senescence stages.

As potatoes, corn, and soybeans have different canopy structures and leaf and stem morphologies, their water absorption capacities vary. Canopies play a key role in biophysical processes by partitioning water fluxes and redistributing energy (Bonan, 2008; Fan et al., 2019). Despite the importance of transpiration through leaf stomata in the terrestrial water cycle, vegetation leaves and stems contribute to intercepting precipitation and partitioning it into water evaporation, runoff, and infiltration at the continental scale (Alton et al., 2009; Schlesinger & Jasechko, 2014). Consequently, constraining canopy surface resistance to realistic values reduces monthly ET bias. Plant functional type differences in canopy structure and growing season length were largely responsible for species' differences in transpiration per unit canopy area, according to Peters et al. (2010). Water loss from crops can also be significantly affected by management practices, such as irrigation, fertilization, and canopy shading (Sharma & Irmak, 2022). Additionally, the current ET model must be improved to represent energy interactions better and incorporate these key processes. Priestley-Taylor has, however, the advantage of being easier to implement than other more complex and computationally expensive methods.

Almost all studies of energy balance and the interaction between the land surface and the atmosphere, including surface ET, require estimation of LW radiation. More accurate sensors and EC towers are critical for studies in energy-limited areas, which can provide additional insight into

the relevance of ET processes. Multiple field measurements across the WCS contribute to the robustness of our optimization approach. The WISP workflow corrected here can continuously improve optimization accuracy by taking advantage of the increasing availability of crop ET measurements with the increased deployment of EC towers in the future. Our data-model integration framework is generalizable to other regions despite the Priestley-Taylor parameters being tailored to WCS's cropland. In addition to the WCS, the Priestley-Taylor approach can be used to assess daily ET and water use in multiple regions once calibrated and tested with local field data.

2.5 Conclusion

It is becoming increasingly important for water management to estimate crop evapotranspiration from the field to the watershed scale, especially when irrigated agriculture is used. An approach to treating net longwave radiation and correcting ET models that fail to account for energy interaction is presented in this paper. Due to large systematic underestimation, the default WISP has large biases (both in sign and magnitude) in representing LW_{net} radiation that propagate to affect ET_a estimation. Especially for energy-limited regions, where ET tracks net radiation, the improved parameterization has significant implications for regional surface energy budgets. Our goal was to demonstrate how to simultaneously calibrate and reduce errors in regional WISP ET_a model parameters. To achieve this, field measurements from a densely irrigated area of WCS were used to calibrate the WISP model. From our analysis and results, we concluded that clear sky atmospheric emissivity ($\varepsilon_{a,clr}$) corrections provide an accurate estimate of the LW_{net} and can be used to estimate ET_a precisely. The calibrated model

reduced the ET_a simulation error compared to the Ameriflux EC tower observations. Observed ET_a for training sites ranged from 0.5-5.7 mm d^{-1} (mean 2.9 mm d^{-1}) for potatoes at our training sites. WISP overestimated ET_a with MAE of 1.15 mm d^{-1} and NSE of -0.08. However, after correction, the obvious agreement can be seen between WISP ET_a and observed ET_a with reduced MAE of 0.72 mm d^{-1} and NSE of 0.52. Moreover, the model performance was mainly consistent for validation sites and crops.

Our finding has significant implications for uncertainty studies of ET and hydrologic models. In addition to evaluating current ET and PET models, it would also be desirable to incorporate energy corrections of latent fluxes into the model for explicit consideration of input energy uncertainty. Even though flux measurements are sparse, they can be skillfully used for model evaluations when in situ meteorological and radiative measurements are available. As a result, the methodology presented here can be used to estimate latent heat fluxes to evaluate models or estimates of annual ET to analyze water budgets. It is demonstrated in this study that remotely sensed evapotranspiration estimates could be used in conjunction with EC observations to inform water resources planning and management. For local water budget estimation of crop water use variation within WCS, we provided more realistic estimates of clear-sky atmospheric emissivity parameters, longwave net radiation and evapotranspiration, emphasizing the importance of crop planning.

2.6 Acknowledgements

We acknowledge support from the Wisconsin Vegetable and Potato Growers Association (WPVGA), the Wisconsin Department of Natural Resources, and the UW Center for Climatic

Research Climate, People, and Environment Program, and the UW Research Forward program, and thank J Thom, T Houlihan, J Pavelski for site support at US-CS1 and US-CS3-US-CS6. S Brevert, A Khan, and S Wiesner assisted with US-CC1 and US-CC2 for which we are grateful from support from Flyte Family Farms. DR and JC are supported by the NASA Carbon Cycle & Ecosystems program (NNX17AE16G)

2.7 Bibliography

AghaKouchak, A., Cheng, L., Mazdiyasni, O., Farahmand, A., 2014. Global warming and changes in risk of concurrent climate extremes: Insights from the 2014 California drought. Geophys. Res. Lett. 41. https://doi.org/10.1002/2014GL062308

Aguiar, L.J.G., Costa, J.M.N. da, Fischer, G.R., Aguiar, R.G., Costa, A.C.L. da, Ferreira, W.P.M., 2011. Estimativa da radiação de onda longa atmosférica em áreas de floresta e de pastagem no sudoeste da Amazônia. Rev. Bras. Meteorol. 26. https://doi.org/10.1590/s0102-77862011000200006

Alados-Arboledas, L., Jiménez, J.I., Castro-Díez, Y., 1986. Thermal radiation from cloudless skies in Granada. Theor. Appl. Climatol. 37. https://doi.org/10.1007/BF00866107

Alison Duff, Ankur Desai (2020), AmeriFlux BASE US-DFC US Dairy Forage Research Center, Prairie du Sac, Ver. 1-5, AmeriFlux AMP, (Dataset). https://doi.org/10.17190/AMF/1660340

Alison Duff, Ankur Desai, Valentin Picasso Risso (2021), AmeriFlux BASE US-DFK Dairy Forage Research Center - Kernza, Ver. 1-5, AmeriFlux AMP, (Dataset). https://doi.org/10.17190/AMF/1825937

Allen, R.G., Pereira, L.S., Howell, T.A., Jensen, M.E., 2011. Evapotranspiration information reporting: I. Factors governing measurement accuracy. Agric. Water Manag. https://doi.org/10.1016/j.agwat.2010.12.015

Althoff, D., Dias, S.H.B., Filgueiras, R., Rodrigues, L.N., 2020. ETo-Brazil: A Daily Gridded Reference Evapotranspiration Data Set for Brazil (2000–2018). Water Resour. Res. 56. https://doi.org/10.1029/2020WR027562

Alton, P., Fisher, R., Los, S., Williams, M., 2009. Simulations of global evapotranspiration using semiempirical and mechanistic schemes of plant hydrology. Global Biogeochem. Cycles 23. https://doi.org/10.1029/2009GB003540

Anar, M.J., Lin, Z., Hoogenboom, G., Shelia, V., Batchelor, W.D., Teboh, J.M., Ostlie, M., Schatz, B.G., Khan, M., 2019. Modeling growth, development and yield of Sugarbeet using DSSAT. Agric. Syst. 169. https://doi.org/10.1016/j.agsy.2018.11.010

Anderson, M., Gao, F., Knipper, K., Hain, C., Dulaney, W., Baldocchi, D., Eichelmann, E., Hemes, K., Yang, Y., Medellin-Azuara, J., Kustas, W., 2018. Field-scale assessment of land and water use change over the California delta using remote sensing. Remote Sens. 10. https://doi.org/10.3390/rs10060889

Anderson, M.C., Allen, R.G., Morse, A., Kustas, W.P., 2012. Use of Landsat thermal imagery in monitoring evapotranspiration and managing water resources. Remote Sens. Environ. 122. https://doi.org/10.1016/j.rse.2011.08.025

Ångström, A. (1915). A Study of the Radiation of the Atmosphere: Based upon Observations of the Nocturnal Radiation during Expeditions to Algeria and to California (Vol. 65). Smithsonian Institution.

Ankur Desai (2021), AmeriFlux BASE US-CS1 Central Sands Irrigated Agricultural Field, Ver. 2-5, AmeriFlux AMP, (Dataset). https://doi.org/10.17190/AMF/1617710

Ankur Desai (2021), AmeriFlux BASE US-CS3 Central Sands Irrigated Agricultural Field, Ver. 3-5, AmeriFlux AMP, (Dataset). https://doi.org/10.17190/AMF/1617713

Ankur Desai (2022), AmeriFlux BASE US-CS4 Central Sands Irrigated Agricultural Field, Ver. 3-5, AmeriFlux AMP, (Dataset). https://doi.org/10.17190/AMF/1756417

Ankur Desai (2022), AmeriFlux BASE US-CS5 Central Sands Irrigated Agricultural Field, Ver. 1-5, AmeriFlux AMP, (Dataset). https://doi.org/10.17190/AMF/1846663

Arnold, J.G., Moriasi, D.N., Gassman, P.W., Abbaspour, K.C., White, M.J., Srinivasan, R., Santhi, C., Harmel, R.D., Van Griensven, A., Van Liew, M.W., Kannan, N., Jha, M.K., 2012. SWAT: Model use, calibration, and validation. Trans. ASABE 55

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X., Malhi, Y., Meyers, T., Munger, W., Oechel, W., Paw, U.K.T., Pilegaard, K., Schmid, H.P., Valentini, R., Verma, S., Vesala, T., Wilson, K., Wofsy, S., 2001. FLUXNET: A New Tool to Study the Temporal and Spatial Variability of Ecosystem-Scale Carbon Dioxide, Water Vapor, and Energy Flux Densities. Bull. Am. Meteorol. Soc. https://doi.org/10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2

Baldocchi, D., Knox, S., Dronova, I., Verfaillie, J., Oikawa, P., Sturtevant, C., Matthes, J.H., Detto, M., 2016. The impact of expanding flooded land area on the annual evaporation of rice. Agric. For. Meteorol. 223. https://doi.org/10.1016/j.agrformet.2016.04.001

Bengtsson, L., 2010. The global atmospheric water cycle. Environ. Res. Lett. https://doi.org/10.1088/1748-9326/5/2/025202

Bilbao, J., de Miguel, A.H., 2007. Estimation of daylight downward longwave atmospheric irradiance under clear-sky and all-sky conditions. J. Appl. Meteorol. Climatol. 46. https://doi.org/10.1175/JAM2503.1

Black, T.A., Gardner, W.R., Tanner, C.B., 1970. Water Storage and Drainage under a Row Crop on a Sandy Soil 1 . Agron. J. 62. https://doi.org/10.2134/agronj1970.00021962006200010016x

Blöschl, G., Bierkens, M.F.P., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., et al., 2019. Twenty-three unsolved problems in hydrology (UPH)–a community perspective. Hydrol. Sci. J. 64. https://doi.org/10.1080/02626667.2019.1620507

Bohn, T.J., Vivoni, E.R., 2016. Process-based characterization of evapotranspiration sources over the North American monsoon region. Water Resour. Res. 52. https://doi.org/10.1002/2015WR017934

Bonan, G.B., 2008. Forests and climate change: Forcings, feedbacks, and the climate benefits of forests. Science (80-.). https://doi.org/10.1126/science.1155121

Bradbury, K., Krause, J., Fienen, M., Kniffin, M., 2017. A groundwater flow model for the Little Plover River Basin in Wisconsin's Central Sands, Wisconsin Geological and Natural History Survey.

Brutsaert, W., 1981. Evaporation into the atmosphere. Theory, history, and applications. Evaporation into Atmos. Theory, Hist. Appl. https://doi.org/10.1029/eo063i051p01223-04

Cai, J.B., Liu, Y., Xu, D., Paredes, P., Pereira, L.S., 2009. Simulation of the soil water balance of wheat using daily weather forecast messages to estimate the reference evapotranspiration. Hydrol. Earth Syst. Sci. https://doi.org/10.5194/hess-13-1045-2009

Chen, X., Su, Z., Ma, Y., Trigo, I., Gentine, P., 2021. Remote Sensing of Global Daily Evapotranspiration based on a Surface Energy Balance Method and Reanalysis Data. J. Geophys. Res. Atmos. 126. https://doi.org/10.1029/2020JD032873

Choi, M., Jacobs, J.M., Kustas, W.P., 2008. Assessment of clear and cloudy sky parameterizations for daily downwelling longwave radiation over different land surfaces in Florida, USA. Geophys. Res. Lett. 35. https://doi.org/10.1029/2008GL035731

Cui, Y.J., Gao, Y.B., Ferber, V., 2010. Simulating the water content and temperature changes in an experimental embankment using meteorological data. Eng. Geol. 114. https://doi.org/10.1016/j.enggeo.2010.06.006

de Oliveira, R.G., Valle Júnior, L.C.G., da Silva, J.B., Espíndola, D.A.L.F., Lopes, R.D., Nogueira, J.S., Curado, L.F.A., Rodrigues, T.R., 2021. Temporal trend changes in reference evapotranspiration contrasting different land uses in southern Amazon basin. Agric. Water Manag. 250. https://doi.org/10.1016/j.agwat.2021.106815

Dilley, A.C., O'Brien, D.M., 1998. Estimating downward clear sky long-wave irradiance at the surface from screen temperature and precipitable water. Q. J. R. Meteorol. Soc. 124. https://doi.org/10.1256/smsqj.54902

Djaman, K., O'Neill, M., Owen, C.K., Smeal, D., Koudahe, K., West, M., Allen, S., Lombard, K., Irmak, S., 2018. Crop evapotranspiration, irrigation water requirement and water productivity of maize from meteorological data under semiarid climate. Water (Switzerland) 10. https://doi.org/10.3390/w10040405

Eichelmann, E., Hemes, K.S., Knox, S.H., Oikawa, P.Y., Chamberlain, S.D., Sturtevant, C., Verfaillie, J., Baldocchi, D.D., 2018. The effect of land cover type and structure on evapotranspiration from agricultural and wetland sites in the Sacramento–San Joaquin River Delta, California. Agric. For. Meteorol. 256–257. https://doi.org/10.1016/j.agrformet.2018.03.007

Famiglietti, J.S., 2014. The global groundwater crisis. Nat. Clim. Chang. https://doi.org/10.1038/nclimate2425

Fan, Y., Meijide, A., Lawrence, D.M., Roupsard, O., Carlson, K.M., Chen, H.Y., Röll, A., Niu, F., Knohl, A., 2019. Reconciling Canopy Interception Parameterization and Rainfall Forcing Frequency in the Community Land Model for Simulating Evapotranspiration of Rainforests and Oil Palm Plantations in Indonesia. J. Adv. Model. Earth Syst. 11. https://doi.org/10.1029/2018MS001490

Fisher, J.B., Lee, B., Purdy, A.J., Halverson, G.H., et al., 2020. ECOSTRESS: NASA's Next Generation Mission to Measure Evapotranspiration From the International Space Station. Water Resour. Res. 56. https://doi.org/10.1029/2019WR026058

Fisher, J.B., Melton, F., Middleton, E., Hain, C., Anderson, M., Allen, R., McCabe, M.F., Hook, S., Baldocchi, D., Townsend, P.A., Kilic, A., Tu, K., Miralles, D.D., Perret, J., Lagouarde, J.P., Waliser, D., Purdy, A.J., French, A., Schimel, D., Famiglietti, J.S., Stephens, G., Wood, E.F., 2017. The future of evapotranspiration: Global requirements for ecosystem functioning, carbon and climate feedbacks, agricultural management, and water resources. Water Resour. Res. https://doi.org/10.1002/2016WR020175

Fisher, J.B., Tu, K.P., Baldocchi, D.D., 2008. Global estimates of the land-atmosphere water flux based on monthly AVHRR and ISLSCP-II data, validated at 16 FLUXNET sites. Remote Sens. Environ. 112. https://doi.org/10.1016/j.rse.2007.06.025

Fisher, J.B., Whittaker, R.J., Malhi, Y., 2011. ET come home: Potential evapotranspiration in geographical ecology. Glob. Ecol. Biogeogr. https://doi.org/10.1111/j.1466-8238.2010.00578.x

Flerchinger, G.N., Xaio, W., Marks, D., Sauer, T.J., Yu, Q., 2009. Comparison of algorithms for incoming atmospheric long-wave radiation. Water Resour. Res. 45. https://doi.org/10.1029/2008WR007394

Foley, J.A., Ramankutty, N., Brauman, K.A., Cassidy, E.S., Gerber, J.S., Johnston, M., Mueller, N.D., O'Connell, C., Ray, D.K., West, P.C., Balzer, C., Bennett, E.M., Carpenter, S.R., Hill, J., Monfreda, C., Polasky, S., Rockström, J., Sheehan, J., Siebert, S., Tilman, D., Zaks, D.P.M., 2011. Solutions for a cultivated planet. Nature 478. https://doi.org/10.1038/nature10452

Formetta, G., Bancheri, M., David, O., Rigon, R., 2016. Performance of site-specific parameterizations of longwave radiation. Hydrol. Earth Syst. Sci. 20. https://doi.org/10.5194/hess-20-4641-2016

Giles-Hansen, K., Wei, X., 2021. Improved Regional Scale Dynamic Evapotranspiration Estimation Under Changing Vegetation and Climate. Water Resour. Res. 57. https://doi.org/10.1029/2021WR029832

Glenn, E.P., Nagler, P.L., Huete, A.R., 2010. Vegetation Index Methods for Estimating Evapotranspiration by Remote Sensing. Surv. Geophys. 31. https://doi.org/10.1007/s10712-010-9102-2

Grafton, R.Q., Williams, J., Jiang, Q., 2017. Possible pathways and tensions in the food and water nexus. Earth's Futur. 5. https://doi.org/10.1002/2016EF000506

Guo, Y., Cheng, J., Liang, S., 2019. Comprehensive assessment of parameterization methods for estimating clear-sky surface downward longwave radiation. Theor. Appl. Climatol. 135. https://doi.org/10.1007/s00704-018-2423-7

Hain, C.R., Crow, W.T., Anderson, M.C., Tugrul Yilmaz, M., 2015. Diagnosing neglected soil moisture source-sink processes via a thermal infrared-based two-source energy balance model. J. Hydrometeorol. 16. https://doi.org/10.1175/JHM-D-14-0017.1

Hatfield, J.L., Reginato, R.J., Idso, S.B., 1983. Comparison of long-wave radiation calculation methods over the United States. Water Resour. Res. 19. https://doi.org/10.1029/WR019i001p00285

Haucke, J., Clancy, K., Kraft, G., 2016. Tools to Estimate Groundwater Levels in the Presence of Changes of Precipitation and Pumping. J. Water Resour. Prot. 08. https://doi.org/10.4236/jwarp.2016.812084

Hemmati, S., Gatmiri, B., Cui, Y.J., Vincent, M., 2012. Thermo-hydro-mechanical modelling of soil settlements induced by soil-vegetation-atmosphere interactions. Eng. Geol. 139–140. https://doi.org/10.1016/j.enggeo.2012.04.003

Holmes, T.R.H., Hain, C.R., Crow, W.T., Anderson, M.C., Kustas, W.P., 2018. Microwave implementation of two-source energy balance approach for estimating evapotranspiration. Hydrol. Earth Syst. Sci. 22. https://doi.org/10.5194/hess-22-1351-2018

Howell, T.A., Steiner, J.L., Evett, S.R., Schneider, A.D., Copeland, K.S., Dusek, D.A., Tunick, A., 1993. Radiation balance and soil water evaporation of bare Pullman clay loam soil, in: Management of Irrigation and Drainage Systems : Integrated Perspectives.

Huang, Q., Qin, G., Zhang, Y., Tang, Q., Liu, C., Xia, J., Chiew, F.H.S., Post, D., 2020. Using Remote Sensing Data-Based Hydrological Model Calibrations for Predicting Runoff in Ungauged or Poorly Gauged Catchments. Water Resour. Res. 56. https://doi.org/10.1029/2020WR028205

Hulstrom, R., Bird, R., Riordan, C., 1985. Spectral solar irradiance data sets for selected terrestrial conditions. Sol. Cells 15. https://doi.org/10.1016/0379-6787(85)90052-3

Huntington, T.G., 2006. Evidence for intensification of the global water cycle: Review and synthesis. J. Hydrol. 319. https://doi.org/10.1016/j.jhydrol.2005.07.003

Idso, S.B., 1981. A set of equations for full spectrum and 8- to 14-µm and 10.5- to 12.5-µm thermal radiation from cloudless skies. Water Resour. Res. 17. https://doi.org/10.1029/WR017i002p00295

Introduction to micrometeorology. By S. Pal Arya. Academic Press. Second edition, 2001. xxv + 420 pp. ISBN 0 12 059354 8., 2002. . Introd. to micrometeorology. By S. Pal Arya. Acad. Press. Second Ed. 2001. xxv + 420 pp. ISBN 0 12 059354 8. 128. https://doi.org/10.1256/0035900021643665

Jensen, M.E., Allen, R.G., 2016. Evaporation, evapotranspiration, and irrigation water requirements, Evaporation, Evapotranspiration, and Irrigation Water Requirements. https://doi.org/10.1061/9780784414057

Jiang, L., Islam, S., 2001. Estimation of surface evaporation map over southern Great Plains using remote sensing data. Water Resour. Res. 37. https://doi.org/10.1029/2000WR900255

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S.I., Sheffield, J., et al., 2010. Recent decline in the global land evapotranspiration trend due to limited moisture supply. Nature 467. https://doi.org/10.1038/nature09396

Juszak, I., Pellicciotti, F., 2013. A comparison of parameterizations of incoming longwave radiation over melting glaciers: Model robustness and seasonal variability. J. Geophys. Res. Atmos. 118. https://doi.org/10.1002/jgrd.50277

Khan, M.S., Liaqat, U.W., Baik, J., Choi, M., 2018. Stand-alone uncertainty characterization of GLEAM, GLDAS and MOD16 evapotranspiration products using an extended triple collocation approach. Agric. For. Meteorol. 252. https://doi.org/10.1016/j.agrformet.2018.01.022

Kimball, B.A., Idso, S.B., Aase, J.K., 1982. A model of thermal radiation from partly cloudy and overcast skies. Water Resour. Res. 18. https://doi.org/10.1029/WR018i004p00931

Kittel, C., Nielsen, K., Tøttrup, C., Bauer-Gottwein, P., 2017. Informing a hydrological model of the Ogooué with multi-mission remote sensing data. Geophys. Res. Abstr. 19. https://doi.org/10.5194/hess-2017-549

Kljun, N., Calanca, P., Rotach, M.W., Schmid, H.P., 2015. A simple two-dimensional parameterisation for Flux Footprint Prediction (FFP). Geosci. Model Dev. 8. https://doi.org/10.5194/gmd-8-3695-2015

Koch, J., Zhang, W., Martinsen, G., He, X., Stisen, S., 2020. Estimating Net Irrigation Across the North China Plain Through Dual Modeling of Evapotranspiration. Water Resour. Res. 56. https://doi.org/10.1029/2020WR027413

Kraft, G.J., Clancy, K., Mechenich, D.J., Haucke, J., 2012. Irrigation Effects in the Northern Lake States: Wisconsin Central Sands Revisited. Ground Water 50. https://doi.org/10.1111/j.1745-6584.2011.00836.x

Kucharik, C., Serbin, S., Vavrus, S., Hopkins, E., Motew, M., 2010. Patterns of climate change across Wisconsin from 1950 to 2006. Phys. Geogr. 31. https://doi.org/10.2747/0272-3646.31.1.1

Kunkel, K.E., Karl, T.R., Easterling, D.R., Redmond, K., Young, J., Yin, X., Hennon, P., 2013. Probable maximum precipitation and climate change. Geophys. Res. Lett. 40. https://doi.org/10.1002/grl.50334

Liu, Y., Qiu, G., Zhang, H., Yang, Y., Zhang, Y., Wang, Q., Zhao, W., Jia, L., Ji, X., Xiong, Y., Yan, C., Ma, N., Han, S., Cui, Y., 2022. Shifting from homogeneous to heterogeneous surfaces in estimating terrestrial evapotranspiration: Review and perspectives. Sci. China Earth Sci. https://doi.org/10.1007/s11430-020-9834-y

Luczaj, J., Masarik, K., 2015. Groundwater quantity and quality issues in a water-rich region: Examples from Wisconsin, USA. Resources. https://doi.org/10.3390/resources4020323

Ma, N., Szilagyi, J., Zhang, Y., 2021. Calibration-Free Complementary Relationship Estimates Terrestrial Evapotranspiration Globally. Water Resour. Res. 57. https://doi.org/10.1029/2021WR029691

Mao, J., Fu, W., Shi, X., Ricciuto, D.M., Fisher, J.B., Dickinson, R.E., et al., 2015. Disentangling climatic and anthropogenic controls on global terrestrial evapotranspiration trends. Environ. Res. Lett. 10. https://doi.org/10.1088/1748-9326/10/9/094008

McDermid, S.S., Mahmood, R., Hayes, M.J., Bell, J.E., Lieberman, Z., 2021. Minimizing trade-offs for sustainable irrigation. Nat. Geosci. https://doi.org/10.1038/s41561-021-00830-0

Mecikalski, J.R., Diak, G.R., Anderson, M.C., Norman, J.M., 1999. Estimating fluxes on continental scales using remotely sensed data in an atmospheric-land exchange model. J. Appl. Meteorol. 38. https://doi.org/10.1175/1520-0450(1999)038<1352:EFOCSU>2.0.CO;2

Melo, D.C.D., Anache, J.A.A., Borges, V.P., Miralles, D.G., Martens, B., Fisher, J.B., et al., 2021. Are Remote Sensing Evapotranspiration Models Reliable Across South American Ecoregions? Water Resour. Res. 57. https://doi.org/10.1029/2020WR028752

Mokhtar, A., He, H., Alsafadi, K., Li, Y., Zhao, H., Keo, S., Bai, C., Abuarab, M., Zhang, C., Elbagoury, K., Wang, J., He, Q., 2020. Evapotranspiration as a response to climate variability and ecosystem changes in southwest, China. Environ. Earth Sci. 79. https://doi.org/10.1007/s12665-020-09007-1

Monteith, J.L., 1965. Evaporation and environment. Symp. Soc. Exp. Biol.

Motew, M.M., Kucharik, C.J., 2013. Climate-induced changes in biome distribution, NPP, and hydrology in the Upper Midwest U.S.: A case study for potential vegetation. J. Geophys. Res. Biogeosciences 118. https://doi.org/10.1002/jgrg.20025

Nocco, M.A., Kraft, G.J., Loheide, S.P., Kucharik, C.J., 2018. Drivers of Potential Recharge from Irrigated Agroecosystems in the Wisconsin Central Sands. Vadose Zo. J. 17. https://doi.org/10.2136/vzj2017.01.0008

Ortega-Farias, S., Antonioletti, R., Olioso, A., 2000. Net radiation model evaluation at an hourly time step for mediterranean conditions. Agronomie 20. https://doi.org/10.1051/agro:2000116

Orth, R., Destouni, G., 2018. Drought reduces blue-water fluxes more strongly than green-water fluxes in Europe. Nat. Commun. 9. https://doi.org/10.1038/s41467-018-06013-7

Paul Stoy, Sharifa Brevert (2022), AmeriFlux BASE US-CC1 Coloma Corn 1, Ver. 1-5, AmeriFlux AMP, (Dataset). https://doi.org/10.17190/AMF/1865475

Paul Stoy, Sharifa Brevert, Anam Khan (2022), AmeriFlux BASE US-CC2 Coloma Corn 2, Ver. 2-5, AmeriFlux AMP, (Dataset). https://doi.org/10.17190/AMF/1865476

Peters, E.B., Hiller, R. V., McFadden, J.P., 2011. Seasonal contributions of vegetation types to suburban evapotranspiration. J. Geophys. Res. Biogeosciences 116. https://doi.org/10.1029/2010JG001463

Pool, S., Francés, F., Garcia-Prats, A., Pulido-Velazquez, M., Sanchis-Ibor, C., Schirmer, M., Yang, H., Jiménez-Martínez, J., 2021. From Flood to Drip Irrigation Under Climate Change: Impacts on Evapotranspiration and Groundwater Recharge in the Mediterranean Region of Valencia (Spain). Earth's Futur. 9. https://doi.org/10.1029/2020EF001859

Prata, A.J., 1996. A new long-wave formula for estimating downward clear-sky radiation at the surface. Q. J. R. Meteorol. Soc. 122. https://doi.org/10.1256/smsqj.53305

Saito, H., Šimůnek, J., Mohanty, B.P., 2006. Numerical Analysis of Coupled Water, Vapor, and Heat Transport in the Vadose Zone. Vadose Zo. J. 5. https://doi.org/10.2136/vzj2006.0007

Schlesinger, W.H., Jasechko, S., 2014. Transpiration in the global water cycle. Agric. For. Meteorol. 189–190. https://doi.org/10.1016/j.agrformet.2014.01.011

Schwartz, F.W., Liu, G., Yu, Z., 2020. HESS Opinions: The myth of groundwater sustainability in Asia. Hydrol. Earth Syst. Sci. 24. https://doi.org/10.5194/hess-24-489-2020

Senay, G.B., Budde, M., Verdin, J.P., Melesse, A.M., 2007. A coupled remote sensing and simplified surface energy balance approach to estimate actual evapotranspiration from irrigated fields. Sensors 7. https://doi.org/10.3390/s7060979

Sharma, V., Irmak, S., 2022. Leaf and canopy stomatal resistance, aerodynamic resistance, and evapotranspiration of irrigated continuous no-till and disk-till maize. Agron. J. 114. https://doi.org/10.1002/agj2.20979

Smail, B. (2016). Irrigated land use statistics for Wisconsin. Wisconsin Department of Natural Resources Water Use Section: Madison, WI, USA.

Sörensson, A.A., Ruscica, R.C., 2018. Intercomparison and Uncertainty Assessment of Nine Evapotranspiration Estimates Over South America. Water Resour. Res. 54. https://doi.org/10.1002/2017WR021682

Sorooshian, S., Li, J., Hsu, K.L., Gao, X., 2012. Influence of irrigation schemes used in regional climate models on evapotranspiration estimation: Results and comparative studies from California's Central Valley agricultural regions. J. Geophys. Res. Atmos. 117. https://doi.org/10.1029/2011JD016978

Sriwongsitanon, N., Suwawong, T., Thianpopirug, S., Williams, J., Jia, L., Bastiaanssen, W., 2020. Validation of seven global remotely sensed ET products across Thailand using water balance measurements and land use classifications. J. Hydrol. Reg. Stud. 30. https://doi.org/10.1016/j.ejrh.2020.100709

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., Hsu, K.L., 2018. A Review of Global Precipitation Data Sets: Data Sources, Estimation, and Intercomparisons. Rev. Geophys. 56. https://doi.org/10.1002/2017RG00574

Swinbank, W.C., 1963. Long-wave radiation from clear skies. Q. J. R. Meteorol. Soc. 89. https://doi.org/10.1002/qj.49708938105

Talib, A., Desai, A.R., Huang, J., Griffis, T.J., Reed, D.E., Chen, J., 2021. Evaluation of prediction and forecasting models for evapotranspiration of agricultural lands in the Midwest U.S. J. Hydrol. 600. https://doi.org/10.1016/j.jhydrol.2021.126579

Tanner, C.B. and Gardner, W.R., 1974. Relation of Climate to Leaching of Solutes and Pollutants Through Soils: Report to Environmental Data Service NOAA for NOAA Grant No. NG-34-72, for the Period of June 15, 1973 Through June 15, 1974, Submitted to Arnold Hull. Department of Soil Science, University of Wisconsin.

Teuling, A.J., Hirschi, M., Ohmura, A., Wild, M., Reichstein, M., Ciais, P., Buchmann, N., Ammann, C., Montagnani, L., Richardson, A.D., Wohlfahrt, G., Seneviratne, S.I., 2009. A regional perspective on trends in continental evaporation. Geophys. Res. Lett. 36. https://doi.org/10.1029/2008GL036584

Twine, T.E., Kustas, W.P., Norman, J.M., Cook, D.R., Houser, P.R., Meyers, T.P., Prueger, J.H., Starks, P.J., Wesely, M.L., 2000. Correcting eddy-covariance flux underestimates over a grassland. Agric. For. Meteorol. 103. https://doi.org/10.1016/S0168-1923(00)00123-4

Vahmani, P., Jones, A.D., Li, D., 2022. Will Anthropogenic Warming Increase Evapotranspiration? Examining Irrigation Water Demand Implications of Climate Change in California. Earth's Futur. 10. https://doi.org/10.1029/2021EF002221

Wanders, N., Bierkens, M.F.P., de Jong, S.M., de Roo, A., Karssenberg, D., 2014. The benefits of using remotely sensed soil moisture in parameter identification of large-scale hydrological modelss. Water Resour. Res. 50. https://doi.org/10.1002/2013WR014639

Wang, K., Dickinson, R.E., 2012. A review of global terrestrial evapotranspiration: Observation, modeling, climatology, and climatic variability. Rev. Geophys. https://doi.org/10.1029/2011RG000373

Wang, K., Dickinson, R.E., 2013. Global atmospheric downward longwave radiation at the surface from ground-based observations, satellite retrievals, and reanalyses. Rev. Geophys. 51. https://doi.org/10.1002/rog.20009

Wang, K., Liang, S., 2008. An improved method for estimating global evapotranspiration based on satellite determination of surface net radiation, vegetation index, temperature, and soil moisture. J. Hydrometeorol. 9. https://doi.org/10.1175/2007JHM911.1

Weeks, E. P., & Stangland, H. G. 1971. Effects of irrigation on streamflow in the Central Sand Plain of Wisconsin (p. 113). Washington, DC: US Department of the Interior, Geological Survey, Water Resources Division.

Weeks, E.P., Ericson, D.W. and Holt, C.L.R., 1965. Hydrology of the Little Plover River basin, Portage County, Wisconsin, and the effects of water resource development. Washington, DC: US Government Printing Office.

Weiss, A., 1982. An Experimental Study of Net Radiation, Its Components and Prediction 1 . Agron. J. 74. https://doi.org/10.2134/agronj1982.00021962007400050024x

Willmott, C.J., 1981. On the validation of models. Phys. Geogr. https://doi.org/10.1080/02723646.1981.10642213

Wong, A.J., Jin, Y., Medellín-Azuara, J., Paw U, K.T., Kent, E.R., Clay, J.M., Gao, F., Fisher, J.B., Rivera, G., Lee, C.M., Hemes, K.S., Eichelmann, E., Baldocchi, D.D., Hook, S.J., 2021. Multiscale Assessment of Agricultural Consumptive Water Use in California's Central Valley. Water Resour. Res. 57. https://doi.org/10.1029/2020WR028876

Wutzler, T., Lucas-Moffat, A., Migliavacca, M., Knauer, J., Sickel, K., Šigut, L., Menzer, O., Reichstein, M., 2018. Basic and extensible post-processing of eddy covariance flux data with REddyProc. Biogeosciences. https://doi.org/10.5194/bg-15-5015-2018

Xing, W., Wang, W., Shao, Q., Yu, Z., Yang, T., Fu, J., 2016. Periodic fluctuation of reference evapotranspiration during the past five decades: Does Evaporation Paradox really exist in China? Sci. Rep. 6. https://doi.org/10.1038/srep39503

Zaussinger, F., Dorigo, W., Gruber, A., Tarpanelli, A., Filippucci, P., Brocca, L., 2019. Estimating irrigation water use over the contiguous United States by combining satellite and reanalysis soil moisture data. Hydrol. Earth Syst. Sci. 23. https://doi.org/10.5194/hess-23-897-2019

Zeng, Z., Wang, T., Zhou, F., Ciais, P., Mao, J., Shi, X., Piao, S., 2014. A worldwide analysis of spatiotemporal changes in water balance-based evapotranspiration from 1982 to 2009. J. Geophys. Res. 119. https://doi.org/10.1002/2013JD020941

Zhang, C., Long, D., 2021. Estimating Spatially Explicit Irrigation Water Use Based on Remotely Sensed Evapotranspiration and Modeled Root Zone Soil Moisture. Water Resour. Res. 57. https://doi.org/10.1029/2021WR031382

Zhang, X., Ding, N., Han, S., Tang, Q., 2020. Irrigation-Induced Potential Evapotranspiration Decrease in the Heihe River Basin, Northwest China, as Simulated by the WRF Model. J. Geophys. Res. Atmos. 125. https://doi.org/10.1029/2019JD031058

[2]

Chapter 3

3. Evaluation of prediction and forecasting models for evapotranspiration of agricultural lands in the Midwest U.S.

Abstract

Evapotranspiration (ET) prediction and forecasting play a vital role in improving water use in agriculturally intensive areas. Metrological and biophysical predictors that drive ET in managed landscapes have complex nonlinear relationships. Deep learning and data-driven methods have shown promising performance for identifying the dependencies among variables. Here, we evaluated the potentials of random forest (RF) and long short-term memory (LSTM) neural networks to estimate and forecast daily ET for corn, soybeans, and potatoes in diverse agricultural farms during 2003-2019. The modeling framework was applied for nineteen fields where eddy covariance ET and meteorological observations in the Midwest USA for growing season (April-October) is available. In this study, we applied data-driven models (RF and LSTM) with 3 sets of predictors (5, 11, and 16 predictors). Results show that a 16 predictor RF model (RF_16

[2] Chapter 3 has been published in the Journal of hydrology.
Talib, A., Desai, A. R., Huang, J., Griffis, T. J., Reed, D. E., & Chen, J. (2021). Evaluation of prediction and forecasting models for evapotranspiration of agricultural lands in the Midwest U.S. *Journal of Hydrology*, *600*. https://doi.org/10.1016/j.jhydrol.2021.126579

R^2=0.7, Willmott's skill score=0.90) outperformed a process-based land surface model (LSM R^2=0.57, Willmott's skill score=0.86) for predicting daily ET, while LSTM performance was lower (LSTM_16 R^2=0.65, Willmott's skill score=0.89 and LSTM_11 R^2=0.62, Willmott's skill score=0.86) than RF using the same sets of predictors. Vapor pressure and crop coefficients were identified as the most important predictors for irrigated crops, while short wave radiation and enhanced vegetation index were key predictors for non-irrigated crops. For certain crop types, such as corn and soybeans on fine-grained soils (silt loam), a simpler version RF, using only 11 drivers, can provide comparable results (R^2=0.70 vs 0.69 and Willmott's skill score=0.90 vs 0.88). For short-term 3-day ET forecasting, LSTM is more sensitive to uncertainty in ensemble forecast meteorology than RF. ET forecasts were strongly sensitive to forecast uncertainty of vapor pressure. The proposed modeling architecture provides a field-scale, locally calibrated tool for accurate prediction and short-term forecasting of daily ET in areas where in situ ET, metrological, and biophysical data are lacking.

Keywords

Evapotranspiration; Machine learning; Agriculture; Drought; Irrigation; Forecasting

Key Points

- RF models predict field-scale ET more accurately than LSTM and process-based models
- Vapor pressure and crop coefficients are key predictors for irrigated crops
- ET forecasting for non-irrigated crop requires enhanced vegetation index.
- Short-term (3-day) forecasts have lower uncertainty, higher accuracy using RF

3.1 Introduction

Terrestrial water in the biosphere and atmosphere is linked through evapotranspiration (ET) (Donohue et al., 2010, Priestley and Taylor, 1972, Wei et al., 2017). ET is the second-largest term in the global land surface water budget (Barr et al., 2014, Narasimhan and Srinivasan, 2005, Trenberth et al., 2007, Wang and Dickinson, 2012). In order to understand terrestrial ecosystem processes in a changing climate such as flash droughts (Kim et al., 2019; Otkin et al., 2016), water resource management (e.g., irrigation efficiency), it is important to accurately estimate and forecast ET (Allen et al., 1998, Anderson et al., 2011, Shugart, 1998). Hydrological applications geared towards conservation of water resources especially for irrigation require prediction and forecasting of ET as a fundamental component. Hence for sustainable agriculture, an ET prediction and forecasting tool can be useful for farmers and water managers to handle water resource challenges (Djaman et al., 2020, Moratiel et al., 2020, Payero and Irmak, 2013; Perera et al., 2014). Actual ET can be measured directly using eddy covariance (EC) towers (Baldocchi et al., 2001, Barr et al., 2012 and Wilson et al., 2001) but costs, logistics, and measurement scale inhibit regional and long-term studies such as EC and Bowen ratio methods (Rosenberry et al., 2007). Further, ET needs to be assessed across a range of crop varieties and soil/climate types that influence it, requiring many observation sites. Hence there is a need for models that are based on more readily available drivers to predict and forecast ET for broader applications.

Data from satellite sensors have been used in earlier studies to estimate ET over domains of different regional scales such as watershed or continent (Anderson et al., 2021, Crosbie et al., 2015, Filgueiras et al., 2020, Fishe et al., 2020, Scott et al., 2008 and Yao et al., 2014), though

satellites are hampered by tradeoff in spatial resolution and revisit frequency, cloud cover, and model assumptions used in linking observations of surface reflectance or brightness temperatures to ET. In addition, data assimilation methods (Meng et al., 2009, Xu et al., 2018; Zou et al., 2017) as well as land surface models (Lian et al., 2018 and Vinukollu et al, 2012) have been used. However, the relative error range for ET estimates compared with ground measurements is from 14% to 44% (Long et al., 2014 and Velpuri et al., 2013) due to factors such as spatial variation, heterogeneity, model parametrization, and unconstrained water balance. In addition, while there are many studies to estimate or predict and forecast reference ET in different climatic conditions (e.g., Fang et al., 2018, Kimball et al., 2019, Li et al., 2016, Perera et al., 2014, Yang et al., 2016; Zhao et al., 2019a) there are not many studies for forecasting actual ET in intensively irrigated and non-irrigated areas.

Field-scale crop models are another avenue for predicting ET. Current crop models that are designed to simulate agricultural practices such as soil composition, nutrients, tilling practices, and irrigation scheduling can be coupled with computational hydrologic and land-atmosphere models (Pauwels et al., 2007). The development of these physically-based and spatially explicit representation of land surface interaction and agricultural processes at the farm scale have high computational costs (Chaney et al., 2016 and Clark et al., 2017), which requires significant parameterization and tuning, subject to collection of a myriad of trait and driver datasets. Even though those models accurately simulate hydrological processes, challenges in calibrating these biophysically-based models make accurate physical process simulations at individual fields challenging. In addition, the available data for calibration and validation of these

models, e.g., three-dimensional information about sub-surface heterogeneity (such as soil texture, moisture, and groundwater flow) limit the application of those models for larger areas with intensive agriculture. However, these models are useful for small-scale regional studies.

In addition to process-based hydrological models, empirical models based on statistical correlations of potential evapotranspiration with meteorological parameters have also been used (Valipour et al., 2017). Often, variables like canopy cover is used in these methods to convert potential evapotranspiration to actual evapotranspiration. The problem with such an approach is that performance may significantly depend on the estimate of canopy cover. An alternate approach to existing empirical and physical based methods is to use data-driven methods to estimate actual evapotranspiration.

A variety of data-driven models have been used in ET simulation studies (Deo and Şahin, 2015, Fang et al., 2018, Izadifar and Elshorbagy, 2010, Pandey et al., 2017). It is efficient to combine information from readily available predictors from remote sensing along with ground observation by applying machine learning (ML) methods that may be able to predict and forecast ET based on relationships between input predictors without utilizing field-based physical parameters. ML algorithms extract non-linear relationships hidden in time series or spatial data and then apply those patterns to estimate and forecast future scenarios. For example, Yang et al., (2006) and Tabari et al., (2012) used a support vector machine (SVM) approach to estimate eight-day averaged ET and reference ET respectively using ground observation and remote sensing predictors. Landeras et al., (2009) used autoregressive models to forecast weekly reference ET and Bodesheim et al., (2018) applied a regression trees based random forest (RF) approach for ET

estimation. Without explicit training, RF can manage high dimensional regression problems and extract the interaction among model predictors (Auret and Aldrich, 2012, te Beest et al., 2017). Shiri et al. (2018) used a coupled wavelet-random forest model for estimating reference ET and showed the potential of a tree-based model in terms of the accuracy of the reference ET model. The use of ensemble trees and randomization makes this approach more flexible, simple, robust and avoids overfitting by making the best use of limited data and reliable performance on both training and test data (Zhang et al., 2017 and Chen et al., 2020).

In addition to ensemble trees algorithms, the artificial neural network (ANN) approaches have been used for both reference and actual ET prediction (Abdullah et al., 2015, Cobaner, 2011, Feng et al., 2017, Ferreira et al., 2019, Jung et al., 2011 and Kisi et al., 2018). Most of these ANN approaches such as convoluted neural network (CNN) for ET modeling are based on a feed-forward neural network approach where the algorithm is introduced for a single layer (Tavares et al., 2015, and Yassin et al., 2016). However, for time series analysis, one of the drawbacks of feed-forward ANNs is that any information about the sequence of inputs is lost and data pre-processing for singular spectrum analysis of time series in these models require complicated procedures (Sahoo et al., 2017). In addition, traditional ANNs also have a problem of exploding or vanishing gradient (Rangapuram et al., 2018). Hence a special type of neural network architecture, recurrent neural networks (RNNs) is designed where input is processed in its sequential order to understand temporal dynamics (Carriere et al., 1996). For problems such as ET prediction and time series forecasting, for which order of the input variables is important, a specific kind of RNN is Long Short-Term Memory (LSTM) that can solve the problem of vanishing gradient. Since our

study focuses on time series prediction and forecasting, RNN such as LSTM along with ensemble trees algorithm such as RF is a suitable choice.

In LSTM, connections between units and cells allow data to move in a forward and backward direction within the model framework. This method helps to overcome the problem of learning lagged dependencies found in traditional RNN. In the case of the water cycle, such an approach allows the model to preserve previous information for future uses such as water storage effects (e.g. snow) or shallow groundwater-driven systems. Kao et al. (2020) used an LSTM model to forecast floods in inundation-prone areas and found that LSTM can be used to link the sequence of rainfall with a sequence of runoff. In addition, Kratzert et al. (2019) applied process-based constraints on an LSTM modeling framework to simulate runoff for a variety of watersheds and found that LSTM outperformed benchmark physically-based coupled models.

As noted above, challenges in existing methods for predicting and forecasting actual ET are the need for extensive parametrization, lack of relevant data drivers, the computational cost of process-based models, and lack of direct estimate of actual ET from empirical models. Knowledge of the performance of data-driven models in different types of irrigated and non-irrigated crops under different soil types is still partial and fragmented. In addition, models in existing studies have only been applied on limited test data sets. Few studies have evaluated the relative contributions of the different input datasets (predictors) to the accuracy and uncertainty of the actual ET models in agricultural fields, particularly across different management (irrigated vs. rain-fed), crop types, and soil textures.

Here, we ask 1) *how well can empirical ML models predict and forecast ET 3 days in advance in irrigated and rain-fed agricultural lands across the Midwest US*? 2) *what are important drivers for predicting and forecasting ET 3 days in advance in irrigated and non-irrigated areas*? We evaluate two different ML models, RF and LSTM, with differing numbers of predictors (5, 11, 16) across a range of crop and soil texture types where eddy covariance observations were available between 2003-2019. The results of this evaluation allow us to better understand the predictors of accuracy and uncertainty in the ET models and propose a multistep prediction and forecast agricultural ET model that can be applied to locations with limited *in situ* observations. Since there is no clear understanding of minimum required predictors for accurate estimates of ET, our models with different sets of predictors (5,11, 16) can help to understand the need for important or minimum drivers for different crop fields on various soil textures in areas with scarce data.

3.2 Methods

In this paper prediction and forecasting models based on RF and LSTM framework are proposed. For ET prediction, RF and LSTM model with 5, 11, and 16 predictors are proposed. For all model experiments, simulations are based on data from 2003-2019.

3.2.1 Data description

The proposed model performance was assessed by using the observed ET data obtained from the AmeriFlux database or site investigators (Table 3.1) for 19 sites located in the agricultural areas of the US Midwest in states of Iowa, Illinois, Michigan, Minnesota, Nebraska, Ohio and Wisconsin **(Fig. 3.1).** Out of those 19 sites, five are irrigated and 14 are rainfed (Table 3.1). Study sites were all located in a temperate climate with cool to cold winters and hot, humid summers. The dominant crops in those regions are soybeans, potatoes, and corn with coarse-grained (sandy loam, loamy sand, loam) and fine-grained (silt Loam and silt clay) soils. The data duration used during this study ranged from 2003-2019 with a daily time step for continuous variables. After removing outliers, only months with less than 3 days gap were used and years with more than one month of missing data were removed. Data gaps for quality-controlled half-hourly ET observations were filled with post-processing software REddyProc (Wutzler et al. 2018). REddyProc method uses co-variation and temporal autocorrelation of turbulent fluxes and gaps are filled based on available information about air temperature, incoming solar radiation, and vapor pressure deficient based on marginal distribution sampling. Additional meteorological data were obtained from Daymet (Thornton et al., 2014) and North American Land Data Assimilation System (NLDAS) Land Surface Model (LSM) (Xia et al., 2012). In addition, MODIS (Aqua MODIS MYD09GA. Aqua MODIS MYD09GA) satellite data (Vermote, 2015) was also used for enhanced vegetation index (EVI), albedo, and solar zenith angle. Table 3.1 describes the study site locations, duration of measurements, and ancillary information. Summary statistics such as mean, maximum, and variance of ET across different observation sites is included in Table 2.

The selection of model input predictors was due to their influence on ET and their availability for agricultural sites (Fig. 3.2). Sixteen model predictors used on daily time stamp for model predictions include moving average precipitation for 7 days (Prcp7), 15 days (Prcp15), and 30 days (Prcp30), as proxies for soil moisture (because direct soil moisture data was not present at all sites), maximum air temperature (Tmax), long-wave radiation (LW), incoming short-wave radiation (SW), solar zenith angle (SolarZenith), albedo (Albedo), enhanced vegetation index (EVI), soil texture (Soil), irrigated versus non irrigated proxy (Irr_nonirr), crop cover (Crop_cover), crop coefficient (Crop coeff), cumulative growing degree days (CumGDD), wind speed (Wind) and vapor pressure (VP). For RF_5 and LSTM_5 daily air temperature (Tavg) was used while for RF_11, RF_16 and LSTM_11 and LSTM_16 maximum air temperature (Tmax) was used. Since RF_5 and LSTM_5 were based on drivers from Priestley Taylor equation, Tavg was used instead of Tmax or Tmin for simpler models. These predictors were chosen because of their ability to explain physical processes (Cobaner 2011, FAO, 2015, Feng et al 2017) of ET as well as easy availability in most regions. The data source for 16 model predictors along with different commination for predictors for various model versions is included in Table 3.1 and Table 3.3.

Cumulative growing degree days (CumGDD) are associated with different phases of plant development (Cleland et al., 2007) and calculated for all growing seasons based on the method described in Anandhi (2016). Crop coefficients were calculated based on the Food and Agriculture Organization of the United Nations (the FAO-56 method) first proposed by Allen et al., (1998). FAO-56 method provides both transpiration and evaporation from soil and reference ET is

calculated based on Penman–Monteith equation. Based on the related version of FAO-56 method (Allen et al., 1998), adjustments were made according to local crop physical condition.

3.2.2 Random Forest model framework

RF is an ensemble of different trees where trees are built to explain the variability of the output by grouping data in homogenous sets. Unique trees are built by data splitting in random sets with replacement like bootstrapping as well as by random subsets of predictors, which helps to increase diversity among trees (Breiman, 2001).

$$\{h(x, \Theta_t), t = 1,2,3,4 \dots \dots T\}$$

where daily ET (independent variable) is represented by x, T is the number of distinct regression trees and predicted value of regression tree in form of ET is represented by $h(x, \vartheta_t)$. Hence random forest builds a large forest where each tree predicts a value for ET. In this study regression, RF of daily ET is affected by different predictors and the average of all those values is the final prediction of RF.

$$h(x) = \frac{1}{\mathrm{T}} \sum_{t=1}^{T} \left(h(x, \Theta_t)\right)$$

Out-of-bag (oob) sampling is used for RF internal validation. In addition, the importance of each predictor can be determined by holding some predictors constant, while permuting each predictor at a time and then comparing the oob error. The paraments that are tuned during RF calibration include n_estimator (number of trees in the forest), and min_samples_split (minimum number of samples required to split an internal node), and min_samples_leaf (minimum number of samples required to be at a leaf node). The mean of yearly and monthly observed ET,

precipitation, and air temperature was computed across various sites and then sites were split between training and testing dataset such a way that each data set has dry, wet, and average years for representation of site conditions. Three RF models RF_16, RF_11, and RF_5 were built with 16, 11, and 5 predictors respectively (Table 3.3) with 70% of the data were used for training and 30% of the remaining data were used for evaluation/validation based on the hold-out method.

3.2.3 Long Short-Term Memory Network (LSTM)

LSTM is a special kind of RNN, without the limitation to learn time series dependencies between input and output features. One limitation of traditional RNN is the inability to "remember" a sequence with long lengths (e.g., > 10) (Bengio et al., 1994). However, the LSTM framework retains memory about the previous timestamp which can help to model lags in energy balance fluxes. The information about long-term memory for each time step is contained in cell state or cell memory $\boldsymbol{c}_t$ of LSTM and sequence of inputs (model predictors) as x is presented in the model and output (predicted or forecast ET) is obtained as h while six parameters show in equations below are updated at each time step in each cell.

Feed-Farward ANNs such as CNN does not store information in memory. We compared LSTM with CNN and chose LSTM algorithm for our prediction and forecasting based on performance . All LSTM models outperformed CNN models. For example, NSE and Willmott's skill score for LSTM_16 was 0.65 and 0.88 respectively while NSE and Willmott's skill score for CNN_16 was 0.53 and 0.84 (Fig. S3.1 and Table S3.1 in supplementary materials).

In LSTM model a sigmoid function is computed by a forget gate (f_t) on new input x_t and previous result h_{t-1}. The sigmoid function is a smooth, differentiable nonlinear function that produces non-binary activation where weights can be updated with every data point. The differentiable activation function is necessary because it can compute the gradient which is required for training via backpropagation. In addition, it can be derived from a maximum entropy model. The sigmoid function helps the forget gate to decide what information needs to be remembered and what information can be discarded from memory. The sigmoid function is also provided with adjustable weights (W) and biases (b) in each LSTM cell. The new information that is going to be remembered is placed in a cell state with the help of the input gate (it), which is also calculated by a sigmoid function. A tanh function is used to calculate a new cell state ($\vec{c}_t$). The output gate regulates the information of the state of cell $\boldsymbol{c}_t$ using a sigmoid function.

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (1)$$

$$\vec{c}_t = tanh(W_{\vec{c}} x_t + U_{\vec{c}} h_{t-1} + b_{\vec{c}}) \quad (2)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (3)$$

$$C_t = f_i * c_{t-1} + i_i \vec{c}_t \quad (4)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_0) \quad (5)$$

$$h_t = \tanh(c_t) * o_t \quad (6)$$

where matrices of weights from the input, forget, and output gates to the input are denoted by Wi, Wf, and Wo, respectively. The bias vectors for input, forget, and output gates are shown by bi, bf, bo, respectively. The hidden layer matrix of weights from the input, forget, and output gates

are represented by Ui, Uf, and Uo, respectively. Logistic sigmoid σ is an element-wise non-linear activation function and the element-wise multiplication of two vectors is denoted with *.

In this study, three LSTM models LSTM_16, LSTM_11, LSTM_5 were built with 16, 11 and 5 predictors respectively. Here the model with 16 predictors is assumed to account for more variability than a simple model of 5 predictors because it includes predictors related to both meteorological and biophysical processes. The model with 5 predictors was built to make a model based on inputs from the common physical ET model of Priestley-Taylor, while the model with 11 predictors was used as an intermediate framework between the complex and simple model. Each model had two fully connected layers. The first layer is called the encoder layer with 50 neurons and that layer is responsible for reading and interpreting the input sequence. Initially, the model was run with 25, 50, 100, and 200 neurons, and an optimal number of 50 neurons was selected for layer 1 based on lower ubRMSE and higher Willmott's index (Fig. S3.2 in supplementary materials).

In order to combat the problem of overfitting, a regularization method of "dropout" was applied after the first layer where the dropout value is a percentage between 0 (no dropout) and 1 (no connection) for LSTM units (Kratzert et al., 2018). Models were tested using different values for dropout and evaluation statistics were calculated to find the optimal number of neurons. In addition, training and testing data performance was compared to avoid an overfitting or underfitting problem (Fig. S3.2 in supplementary materials). A dropout value of 0.10 was applied in LSTM_16, and a dropout value of 0.25 was applied for LSTM_11 and LSTM_5 (Table 3.4). After the dropout function, a decoder layer was applied which used the output of the encoder (first

layer) as an input. A second LSTM layer that comes after the encoder had 25, 50, and 100 neurons for LSTM_5, 11, and 16 respectively (Table 3.4). The optimal number of neurons was obtained by using different combinations of neurons and dropout factors until reduced ubRMSE was obtained without overfitting or undefining model.

Lastly, two dense layers were applied. The model was calibrated using ADAM (*adaptive moment estimation)* optimizer and mean squared error loss function. A moderate rate of 0.001 is used for the ADAM optimizer for learning. During the calibration process, it was observed that a high learning rate of 0.1 missed the optimal point ($R^2 > 0.6$) and a smaller learning rate of 10^{-6} led to a longer convergence time for the model (Zhang et al., 2018).

Randomly selected 70% of raw data were used for calibration and 30% of the remaining data were used for evaluation using the hold-out method. During the training and optimization of the learning algorithm, a loss function was used to estimate the error of the current state of the model. The purpose of this loss function is to reduce the loss of the next evaluation by updating weights (Kratzert et al., 2018). During training initial loss function was 0.59, 0.72 and 0.70 for LSTM_16, LSTM_11 and LSTM_5 respectively that was reduced to 0.30, 0.51 and 0.60 for LSTM_16, LSTM_11 and LSTM_5 respectively by end of training.

3.2.4 Land Surface Model

We benchmarked our empirical models against output from process-based model ET from the North American Land Data Assimilation System (NLDAS) version 2 LSM model (Xia et al., 2012). Daily ET data were downloaded from Land Data Assimilation System (LDAS) (https://ldas.gsfc.nasa.gov/nldas/). Penman-Monteith equation is used in NLDAS-Noah LSM

energy balance for latent heat flux here ET is based on evaporation, and plant transpiration is driven by soil moisture stress on the top layer of the soil profile (Chen et al., 1996). Hence under wet conditions, ET is equal to potential evapotranspiration. Richards (Richards, 1931) equation is used in this model to simulate soil moisture dynamics. Root zone plant transpiration is driven by canopy interception and canopy resistance that is parameterized by solar radiation, air temperature, vapor pressure, and soil moisture (Koster and Suarez, 1994). Surface albedo is simulated based on diurnal variations and simulated LAI (500 m resolution) varies seasonally as well as spatially and the minimum stomatal resistance parameters are based on vegetation types. In addition, surface runoff is calculated based on the Simple Water Balance (SWB) model, and baseflow is represented by gravity drainage (Chen and Dudhia, 2001).

3.2.5 Significant predictors

The significance of each predictor variable with respect to its effect on the RF model is displayed by predictor importance. The RF model algorithm calculates predictor importance internally to account for bias in test data (Liaw and Wiener 2002). In decision trees, the node uses predictors to split values of output (ET) and similar values of the output (ET) end up in the same set after the split. Predictor importance is measured by measuring how much each predictor contributes to decreasing the variance. In order words. importance of predictor is based on the frequency of its inclusion in the sample by all trees and it is a measure of how much removing a predictor decreases accuracy (Breiman, 2002 and Pedregosa 2011). A decrease in variance from each predictor is averaged in a forest and predictors are ranked according to this measure. We used the Sklearn algorithm in python 3.7 to calculate the importance score for each predictor

after training and the score is scaled to 1 to calculate the influence of each predictor on ET. Therefore, the sum of the importance of all predictors is equal to one, and the higher the value associated with a predictor, the more important that predictor. The importance of model predictors was calculated for both prediction and forecast model versions of RF.

The LSTM algorithm does not have a built-in variable importance selection criterion. So predictors' importance was measured in terms of change in NSE by removing certain predictors and by comparing the change in NSE with the NSE obtained from the original LSTM_16 model.

3.2.6 Forecast model

After evaluation of ET prediction, we also proposed a multistep forecast model that can forecast ET three days ahead of time using RF and LSTM models as described above **(Fig. 3.2).** At each daily time step, there are three ET forecasts: 1) day 1 ET (tomorrow ET), day 2 ET (ET day after tomorrow), day 3 ET (ET three days from today). Forecasts were made by integrating the uncertainty of forecast meteorology through ensemble simulation. Hence, along with 16 model predictions that were used for the prediction model (Table 3.3), input meteorological predictors from re-forecasts from the National Oceanic and Atmospheric Administration's (NOAA's) National Centers for Environmental Prediction (NCEP) 11-member Global Ensemble Forecast System Reforecast version 2 (GEFSRv2) were propagated into each model to make forecasts (Hamill 2013). The uncertainty in meteorological forecasts of GEFS was quantified by generating ten ensembles of multiple ET forecasts, each perturbed from the original observations (or control). RF_16 and LSTM_16 versions were used for forecasting ET.

RF_16 prediction model and initial forecasting model runs provided us with identification of important variables as described above. We also measured the Pearson correlation between predictors and ET to evaluate forecast reliability. In addition, for LSTM_16 we did some initial model runs with different combinations of predictors and only used those predictors that helped to improve the accuracy of the model (using ubRMSE, MAE, AIC criteria). Hence based on initial model runs, information from the prediction model, and literature review (Fang et al., 2018), only those meteorological predictors were selected that were the main drivers of future ET, i.e., maximum and incoming solar radiation, minimum temperature, and precipitation. So for the day 1 ET forecast, forecast meteorology for the next day was included in the model. For day 2 ET, forecast meteorology of days 1 and 2 were included. For day 3 ET, forecast meteorology of days 1, 2, and 3 were included.

3.2.7 Model evaluations

There are 19 sites with 14 rain-fed sites, and five irrigated sites, with a total of seventeen site-years (growing season April-October) of observations, or 26,331 daily observations of ET. Thirteen of the 19 sites were used for training where for one of the sites 80% of data was used in training and the remaining 20% of data from the same site was used in testing. These thirteen sites were used for training while seven sites were held-out and used exclusively for testing. In total, 70% of observed ET data (18,481 daily ET observations), from the 13 different agricultural sites for corn, soybeans, and potatoes, was used for calibration, and data from the remaining seven agricultural sites were used for evaluation for the time period 2003-2019 (7,850 daily ET observations). To test the accuracy of the calibrated models, a subset of data was used to

determine the optimal number of trees in RF and hidden neurons and layers in LSTM and an optimum or satisfactory point for the calibration without overfitting the models for one set of data.

For statistical analysis, coefficient of determination R^2, Pearson correlation coefficient, *Nash–Sutcliffe* (NSE), Willmott's skill score or index of model performance (Willmott, 1981), mean absolute error (MAE), unbiased root mean square error (ubRMSE), RMSE-observations standard deviation ratio (RSR) (Moriasi et al., 2015), percentage bias (Pbias) were used to assess the predictive ability of the proposed RF and LSTM models. In addition, Akaike's Information Criteria (AIC) metric was also used to see the effect of penalization of additional drivers to the model (*Akaike,1970).* AIC adds penalty by including additional predictors in the model that leads to higher error. Hence a more parsimonious model will have lower AIC.

$$AIC = -2\ln(L) + 2k$$

where L is the likelihood and k is the number of parameters. Likelihood is calculated as the log of mean square error.

3.3 Results

3.3.1 RF versus LSTM prediction model evaluation

Fig. 3.3 illustrated the performance of the two ET prediction algorithms for the test data, which demonstrated the ability of the calibrated models to generalize to unseen ET observations (test data) from eddy covariance flux towers across multiple crop types. The evaluation statistics shown in Fig. 3.3 indicated that there is a good agreement between the predicted and observed

ET values across corn, soybeans, and potatoes. For RF 16 model, R^2 and NSE values for the corn vary from 0.53 to 0.70 (Willmott's score 0.85-0.9) in the testing period and for LSTM_16 the R^2 range was 0.56-0.66 and Willmott's skill score (0.80-0.89). Further, LSTM_16 had less bias for the site with a smaller number of observations (potatoes in loamy sand) compared to RF_16 (Fig. 3.3).

The more complex model required a greater number of neurons for the LSTM hidden layer. The number of neurons for different versions of best-fit LSTM models varies from 25-100 (Table 3.4). For the LSTM_16 model, using more than two layers and more than 100 neurons did not improve the model performance on testing data. The run time for the LSTM_16 model and RF_16 model was ten and two minutes, respectively on an Intel CORE i7 9750H CPU, windows 10 X64 based processor.

Both model outputs products closely follow the seasonal growth of crops (Fig. 3.4). During the shoulder months (i.e. September to next May), ET is lower, and as percentage canopy cover increased in June-August, so did ET. In addition, both observations and models are consistent in showing that during dry years (2006, 2010, and 2012), ET is higher than compared to wet years (2014-2018) across crop types. For example, in the drought of 2012, the ET at US-Ro1 and US-Ro3 was above 6 mm day^{-1} while it was less than 6 mm day^{-1} in the wet summer of 2015.

The consistency of modeled ET against the ground truth differs based on the regional characteristic and amount of data available for calibration. For example, in sites US-CS1 and US-CS3, RF model predictors could well track the dynamics of the water loss caused by an increase in canopy cover. However, LSTM_16 had lower Pbias (-5.7%) but higher error (Fig. 3.3,4) than RF_16 (-24.1% Pbias). In general, RF_16 had a higher bias, but lower error compared to LSTM_16

during months when irrigation and ET are higher (June, July). We also computed the Empirical Cumulative Distribution Function (ECDF) for the evaluation period under different soil conditions, soil moisture (variable precipitation under wet, dry years), and crop types. The ECDFs of RF_16 and LSTM_16 models match closely with the observations. Compared to extreme events, the middle section of ECDFs curves is better represented by models.

3.3.2 Significant predictors

The significance of each predictor variable with respect to its effect on the RF and LSTM models is displayed by predictor variables' importance. Four predictor variables that explained most of the variance in the data include Enhanced vegetation Index (EVI), solar zenith angle, incoming SW radiation, and CumGDD. These four predictors combined explain 62% of the model variance. **Fig. 3.5** also showed the Pearson correlation coefficient between predictors and ET, which is positively correlated with VP and EVI. Since most of the Midwest regions are not moisture limited and have a humid climate with warm summers, we expect to see a high correlation between ET and maximum daily temperature (during the growing season) compared to the correlation between ET and precipitation (i.e., our soil moisture proxy in the form of moving average precipitation). In irrigated fields, NSE was reduced from 0.7 to 0.52 by removing SW and SolarZenith predictors (Fig. 3S in Supplementary materials) in LSTM. In addition, a change in NSE from 0.6 to 0.47 was observed by removing SW and SolarZenith from rainfed or non-irrigated fields. The positive correlation between maximum daily temperature can be seen in the ranking of the CumGDD predictor among the four most important predictors for RF (**Fig. 3.5)**. In contrast, despite the low direct correlation of soil moisture proxy (seven days average

precipitation), it is among the five most important predictors for the RF model (Fig. 3.5). Crop coefficient also improved model performance by explaining the dynamics of canopies (cover fraction, LAI, greenness). Our analysis for RF model showed that VP and crop coefficients were the most important predictors for irrigated crops, while short wave radiation and enhanced vegetation index were key predictors for non-irrigated crops (Fig. 3.6).

3.3.3 Model performances

Different versions of the RF and LSTM models (complex versus simple models) were also evaluated on a daily timestep in comparison with the daily predictions from the mechanistic model – NLDAS-Noah (Table 3.5). Overall, the RF_16 model resulted in an R^2 of 0.7 with a Pbias of -4.7% while the RF_11 model had an R^2 of 0.7 with Pbias of -5.3% (Table 3.5). The NLDAS-Noah model had a 0.57 R^2 with the lowest Pbias of 0.3 **(Fig. 3.7).** The lowest Pbias for NLDAS-Noah was most likely a result of the averaged ET prediction across a larger geographical area. That leads to a wider spread from the mean estimate on the scatter plot with a ubRMSE of 1.1 mm/day and a lower R^2 of 0.57 for the NLDAS-Noah model (Fig. 3.7).

Residuals were obtained for each model time step (daily) by subtracting the observed ET from the predicted ET. A negative residual value showed that the model underestimates ET while a positive residual means that ET is overestimated. The distribution of residuals is the largest for the testing period. Based on residuals, the RF_11 produced the most accurate results in April and June (with 0.02 and -0.01 mm residuals, respectively) while the RF_16 was the most accurate model in September (Fig. 3.8). In July and August, the NLDAS-Noah model prediction was more accurate compared to other models. This could be because mechanistic models such as

NLDAS_Noah has constrained ET by using soil moisture at different depths. If soil moisture storage is significantly variable due to large ET during the mid-growing seasons (July-August), the mechanistic model may outperform empirical models. In shoulder months, since ET is lower, the coupling/interactions between soil moisture and ET is also lower. Overall models residuals were lower for the shoulder months of April, May, September and October and were in the range from 0.003 to 0.1 mm (overestimate of ET) while in peak warm months of June, July, and August, residuals range from -0.2 to -0.6 mm (underestimate of ET).

For the overall evaluation data set, RF_16 outperformed other models with the lowest AIC and R^2 of 0.7. The performance of the RF_11 was similar. RF_5 and LSTM_5 were the simplest version RF and LSTM, respectively, and produced the highest daily ubRMSE of 0.94-1.20 mm. As the model complexity reduced, ubRMSE and AIC error increased for both LSTM and RF and overall RF consistently outperformed LSTM. Supplementary materials include models results from training data (13 sites with 18481 daily ET observations) and testing data (seven sites with 7850 daily ET observations) in irrigated and rain-fed fields and their comparison with benchmark model (Fig. S3.3-Fig. S3.7 in supplementary materials). In addition, evaluation metrics for RF_16 (overall best model) are calculated for daily ET is each year of testing data in Table S3.2 in supplementary materials.

For non-irrigated crops, the predictors that improved RF_16 and RF_11 performance were similar and additional predictors such as soil texture, crop cover, crop coefficients, and cumulative GDD did not significantly improve the model performance of RF_11 and RF_16. However, this was not the case for the irrigated crops. Here, ET prediction was improved by including additional

information related to physical properties of sites (soil types, crop coefficient, cumulative GDD) and relative AIC error reduced from 0.16 to zero (**Fig. 3.9**), making RF_16 the best model for irrigated crops. AIC score and R^2 were also computed for sites with different crops and soil texture. For all crop types, the simplest versions of models such as RF_5 and LSTM_5 (**Fig. 3.10**, Table 3.6) increased ubRMSE and AIC errors. Soybean and corn on fine-grained soils such as silty loams did not show an increase in R^2 or decrease in ubRMSE and AIC in models by including additional 5 parameters in RF 16 and LSTM 16 model. However, corn and soybeans on coarser soil such as loam showed improved performance with additional information about crop planting and harvest dates, cumulative GDD, and crop coefficients.

3.3.4 Forecast model results

The evaluation was performed for the retrospective period of 2003–2019 (Fig. 3.11). For both RF and LSTM, the overall ET estimate was comparable for day 1, day 2, and day 3 ET forecast. It is observed that as lead time increases, uncertainty and error in forecast increases but for proposed RF and LSTM models there was only a slight increase in MAE from 0.74 to 0.75 mm and from 0.75 to 0.80 mm **(Table 3.7).** The MAE for June, July, and August was higher, in concordance with higher variance on GEFS meteorology ensemble forecast spread **(Fig.11).** This bias was more evident in LSTM models where ensembles estimates showed a wider spread from the mean estimate compared to RF. Overall, the RF forecast model produced results with high confidence (small ensemble standard deviation) compared to LSTM. RF was also more precise and less biased than the LSTM, for example, for day 3 ET forecasting MAE= 0.75 vs. 0.8 and Pbias = -4.1% vs. -5.1% (Table 3.7). However, overall, the difference between the results of the two forecasting

models is not significant (p value for two-tailed t-test is 0.2). Based on variable importance for RF forecast models (Fig. 3.11), VP and SolarZenith explained about 32% and 12% variance in the model. Other important model predictors include Crop_coeff, CumGDD, EVI, and SW_Day3. For LSTM removing day 2 and day 3 SW radiation reduced model NSE from 0.56 to 0.49 (Fig. 3.11) for day 3 ET, indicating significance of meteorological forecast predictors.

While the forecasts appear reliable, there are differences in soil type, climate conditions, and irrigation. RF and LSTM were consistent in prediction on sandy or loamy soils, but underpredicted ET on silty loam (Fig. 3.12). The performance of the daily ET forecast model decreases during extreme conditions. **Fig. 3.13** showed that RF outperforms LSTM for ET forecast for day 3 for irrigated crops (RF NSE=0.70 and Willmott's skill score =0.91 vs LSTM NSE=0.67 and Willmott's skill score =0.90, p value 0.0001) and non-irrigated crops (RF NSE= 0.53 and Willmot' skill score of 0.81 versus LSTM NSE= 0.50 and Willmott's skill score of 0.80 for non-irrigated areas p value 0.07). The difference between RF and LSTM model performance was significant for irrigated sites.

Models performance was also tested for extreme events such as as floods and drought years. **Fig. 3.14** showed that for 2012, a dry year with a flash drought, the difference between the model for day 3 ET forecast estimate is larger during days (July, August) with high temperatures and ET. Similarly, for the year 2017, a wet year, the model for day 3 ET forecast overestimated lower values (~1 mm) of ET. These analyses indicate that there are an under-estimation and over-estimation of the forecasted maximum and minimum values, respectively.

3.4 Discussion

3.4.1 Model evaluations

Overall, we found that empirical ML models can accurately and precisely predict ET across a range of crop and soil types in the upper Midwest USA, with R^2 and NSE equal to 0.70 and ubRMSE from 0.75 and 0.89 mm day^{-1} for RF_16 and LSTM_16 respectively. In general, different versions of RF models had higher R^2 and NSE and lower Pbias than the LSTM, except for irrigated potatoes in sandy loam. We suspect that this result is because we had data for only two growing seasons for irrigated potatoes, thus our results support that while RF can be more accurate, LSTM may be more useful when available data for model calibration is smaller. In addition, the prominent soil type for sites with irrigated potatoes (US-CS1 and US-CS3) is loamy sand, which stimulates rapid water movement through coarse grains after precipitation and irrigation. RF_16 could capture this pattern properly during months with high ET and irrigation during months when ET is higher but not during months with moderate or low ET, while LSTM_16 had a larger variance than the bias during such extreme events. This indicates that when irrigation and ET are higher (June and July), RF_16 had a higher bias, but lower error compared to LSTM_16. The high bias for RF_16 for that site is likely because of RF's greater sensitivity to the size of the training sample. The high bias of RF_16 can also be seen in Iowa (corn and soybeans rotation, loamy soil), Minnesota (corn, soybean rotation silty loam), and Michigan (Corn, Sandy loam). Thus, while RF models outperform LSTM for crop ET, they require more training data. Time series analysis of observed and predicted ET values (Fig. 3.4) shows that data-driven models are well trained for

predicting the daily data. Hence errors in reproducing the daily anomalies are smaller when compared to the errors in the seasonal cycle because of their relative amplitude.

For the evaluation period of different versions of RF and LSTM, residuals (simulated-observed) are roughly normally distributed during the growing season. However, negative residuals in range of -0.25 to -0.75 for different versions of models during peak ET months (July, Aug) showed an underestimation of ET. This difference may also be due to errors in the input data from different sources, or complexities that the model cannot explain e.g., more irrigation during the dry year or not capturing fluxes through the root zone of fine-grained and coarse-grained soils. Our current models do not have irrigation data as a predictor so including it in future research can be useful. Our work is consistent with earlier studies on using ML to estimate water cycle elements. For example, Kratzert et al., (2019) used LSTM in an ungauged basin (with an aridity index from 0.22 to 5.20) to estimate stream flow using static predictors (e.g soil, geology, water content, max LAI) and non-static parameters (e.g precipitation, temperature, solar radiation) and found that ML models can be useful to predict information by extracting complex relationships between diverse data under heterogeneous condition.

3.4.1.1 Model complexity

The complexity of an RF tree grows with an increase in the number of trees in the forest as well as the number of training samples. Hence a simple RF tree with small training samples could not account for variability in the potatoes ET. In RF we also limit the number of variables

to split on in each split that can lead to higher bias in each tree especially when the sample size is small.

Among the ML models, we found that the best overall model to be RF_16. We also observed that more than 300 decision trees for the RF model only improved the accuracy of training data but did not show significant improvement in model accuracy on testing data, and instead only made the proposed approach computationally more intensive. The performance of the RF_16 model is comparable to the process-based NLDAS-Noah model and needs relatively fewer parameters and drivers to estimate ET. The inner structure of RF allows the model to explain the non-linear relationships among ET and important predictors such as EVI, solar zenith angle and incoming SW radiation. RF_16 models outperformed other smaller parameter number models for most of the locations except at corn and soybeans with silt loam soil texture and potatoes in sandy soils. At those locations, we found that a simple version of RF (RF_11) performed better at those locations as well as for non-irrigated crops. For these sites, the complex models (with 16 predictors) were overfitting on training data. In other words, for these crop and soil combinations, an overall simpler model was able to learn the appropriate non-linear relationship and memory (in the case of LSTM) between predictors and memory. Hence, we can expect the performance of LSTM and RF to decline when models are trained on drivers beyond the leading predictors of a hydrologic system. Tennant et al., (2020) observed this decline in performance in the LSTM discharge prediction model in snow-dominated catchment when trained on additional predictors.

Another reason for the divergence in model performance among sites may be related to the observation that irrigated crops have high variability in ET e.g. based on summary statistics in Table 2, irrigated crops in US-Ne2 have maximum daily ET at a higher end (e.g., ~9 mm day^{-1}) with sample variance more than 3 mm compared to non-irrigated crops. Although, we did not observe this high variability in irrigated potatoes in US-CS1 and US-CS3 and irrigated corn in MI sites (US-JCK, Jackson 1), because available data were only from wet years of 2018 and 2019. In addition, when water is sufficient or close to sufficient, the importance of additional predictors such as storage capacity (soil texture) and crop phenology (crop coefficients) become stronger and have a critical role in predicting ET. However, this effect is masked when irrigation is not available, or soil water storage is relatively low in non-irrigated crops (Seneviratne et al., 2010).

When predictors were reduced to only 5, both RF and LSTM performance contained large errors, limiting their utility. This outcome showed the importance of wind speed, solar zenith angle, maximum temperature, albedo, and 30 days average precipitation (as soil moisture proxy) that were excluded in the RF_5 and LSTM_5. Oliveira et al., (2018) also noted that surface energy fluxes that drive ET depend on rainfall and soil moisture, and albedo's influence on net radiation estimates. Thus, we argue that our 11-parameter model is the baseline minimum inputs required to predict ET across a range of crop, soil, and irrigation types. This also suggests that a number of predictors lower than 11 could not explain the variance in ET and it is possible for some sites to build more robust models with 11 predictors instead of 16. However, it's worth noting that the improvement of performance in LSTM and RF is not just from more parameters, but also the more complex models include a greater number of hidden neurons, in the case of LSTM, or decision

trees subsets and nodes in the case of RF. The additional elements provide an additional benefit over easily implementable regression-based models that cannot account for the non-linear interactions among the predictors (e.g. temperature) and ET. For example, Chen et al. (2020a) found that temperature and humidity-based ML models (RF and LSTM) outperformed temperature and humidity based empirical models in areas with limited meteorological data.

Compared to other techniques for ET estimation, the advantage of the proposed ML modeling approach is that these models monitor ET by using fewer parameters and do not rely on the accurate parameterization of mechanistic models or collections of labor-intensive field-scale data (e.g. field-scale leaf area index). However, care must be taken in appropriate model selection because the models are location-dependent and require sufficient calibration and testing data. For example, for soybeans and corn in silt loam, corn with sandy loam soil texture, and potatoes with loamy and sandy texture a comparable level of ET prediction performance can be achieved without using additional parameters about crop coefficients, crop cover, or CumGDD. Hence, ET can be predicted by the readily available biophysical predictors for such locations, in contrast to ET prediction for corn and soybeans with loam soil texture, where model performance is improved by including those biophysical parameters. The importance of cumGDD in daily ET prediction is encouraging as it is readily derived from low-frequency temperature observations and more readily available across more sites than soil moisture.

We found good performance using the same crop coefficients for irrigated and non-irrigated crops. Depending upon the objective and availability of data for a study, different models can be built for a specific crop type and soil texture at a daily time step.

3.4.2 Significant predictors

The predictors importance of the RF model (Figure 5) highlights driving predictors and combats with the black-box nature of some ML models. Our study showed that EVI, solar zenith angle, incoming SW radiation, and cumulative growing degree days are important predictors to predict daily ET for the growing season (April-October) in the Midwest. Similarly, studies based on empirical models (Priestley and Taylor, 1972 and Jensen et al. 1990) and data-driven ML framework (Chen et al., 2020a) evaluated that most of the variation in reference ET can be explained by solar radiation. This result is consistent with our study where incoming solar radiation explaind about 10%-20% variance for irrigated and non-irrigated crops. However, in our study, an additional variation of about 20% was explained by other variables such as EVI and crop coefficient. Zhao et al., (2018) also found that crop coefficients not only correlate with canopy development but also controls seasonal ET partitioning and surface soil moisture. This shows the importance of variable crop coefficients and EVI in predicting ET.

Noting that LSTM_5 and RF_5 residuals were high especially during peak ET months models suggests that wind speed, albedo, and EVI are leading factors that promote enhanced ET. For example, the potential of plants to extract water from soil depth varies during different stages of crop growth, so we can surmise through the lower residuals that it was captured by 11 and 16 predictors model versions. EVI has been used for agricultural drought monitoring (Song and Ma 2011) and the results of this study also suggest the potential of EVI and ET as good indicators of short-term and long-term drought.

Our work is consistent with earlier studies to estimate ET. For example, Cobaner (2011) used fuzzy inference system-based grid partition to estimate reference ET in the moderate Mediterranean climate of California and found that solar radiation, air temperature, and relative humidity as important drivers for ET prediction. Our model was built to estimate daily actual ET for agricultural lands and found that EVI (non-irrigated crops), crop coefficients, and VP (irrigated crops) to be better predictors than solar radiation in Midwest humid-temperate climate.

Feng et al., (2017) used temperature-based RF and generalized regression neural networks (GRNN) to estimate reference ET and found that RF outperformed GRNN. They also noted that without using solar radiation temperature-based RF and GRNN underestimated reference ET. Our model also found that incoming SW radiation was a more important driver than Tmax for actual ET. Wall et al., (2020) used RNN model based on ReLU and sigmoid activation function to estimate actual ET and found that without net radiation, model performance goes down. In our study, we found incoming SW radiation explained higher variance for ET compared to LW radiation, and thus net radiation could be omitted. In terms of RF and LSTM comparison for other hydrological variables, such as snowfall retrievals from microwave humidity sounders, Adhikari et al., (2020) found that RF is more robust than LSTM.

Chen et al. (2020b), which developed an LSTM based actual ET prediction model irrigated maize/corn, found that leaf area index, relative humidity, and solar radiation as important features that drive corn dynamics in a continental monsoon climate. Those predictors are in agreement with physical processes that can affect corn ET. We also found that VP and crop coefficients were more important predictors for irrigated crops compared to non-irrigated, while

incoming SW radiation explained more variation in non-irrigated compared to irrigated crops. Irrigation influences surface temperature, convection, cloud formation (Lohar and Pal 1995), and humidity (Jianping et al. 2002). In irrigated crops, additional water vapor (Boucher et al., 2004) in the atmosphere due to evaporation of irrigated water can explain why vapor pressure is an important driver for irrigated crops, while less surface cooling in non-irrigated land can make incoming SW radiation important driver for those sites.

Since our study had shown that EVI is the most important variable for rain-fed crops, the uncertainty of EVI and associated parameters used in other models (e.g. for deriving leaf area index, LAI) will greatly affect ET estimation/mapping across the globe and improvement in estimating LAI can improve hydrologic and land surface models for ET mapping. Thus, methods to reduce uncertainty in EVI can improve remote sensing estimate of ET (Sharma et al., 2015).

We also found that soil texture is important in improving ET estimation in irrigated fields, which suggests the use of soil texture maps for ET estimation in ET mechanistic models in addition to soil moisture as a limiting factor. Dong et al., (2020) showed that soil moisture and ET coupling strength bias is caused by oversimplification of soil texture effects on soil evaporation stress. A data-driven based hydrodynamic prediction model can benefit from data sets of appropriate temporal and spatial coverage, readily available meteorological, biophysical variables, and advanced RNN such as LSTM (Kratzert et al., 2019**)** as well as robust simple ensemble tree-based RF algorithms.

3.4.3 Forecast models evaluations

We found that RF and LSTM framework can be used for forecasting for three days in advance using gridded forecast meteorology. Based on our hindcast analysis, the RF forecast model provided higher accuracy overall than LSTM, consistent with prediction model evaluation. LSTM forecast model was more sensitive to GEFS meteorology ensembles, where a higher spread from mean forecast ET was observed compared to the RF forecast model and RF can handle multivariate dimensionality (Belgiu and Drăgut, 2016) better than RNN.

ML-based actual ET forecasts are a novel contribution of our research here and demonstrate significant performance across multiple irrigated and non-irrigated crops and soil texture. Short-term ET forecasts have value for irrigation planning considering under-irrigation and over-irrigation can be detrimental for crops and local water supply quantity and quality. We find that vapor pressure, solar zenith angle, and third-day forecasted incoming SW radiation are important predictors for accurate and precise ET forecasts. Ferreira and Cunha (2020) used similar meteorological predictors (maximum air, solar radiation) for multistep forecasting of reference ET and found that deep learning models such as LSTM performed better than classic machine learning models. This is because LSTM process input in its sequential order and overcomes the problem of learning lagged dependencies. In addition, connections between neurons, that allow data to move in forward and backward direction within the modeling framework of LSTM and helps to learn temporal dependencies. Perera et al. (2014) used numeric prediction output for reference ET forecast in Australia and found that forecasting based on air and dew point temperatures leads to better performance for all lead times compared to incoming SW radiation

and attributed the poor performance of incoming SW radiation to error forecast weather meteorology. Our study found incoming SW radiation (forecast) a more important predictor compared to day air temperature for actual ET forecast at all lead time. Higher ET during dry seasons showed that water was not limited due to irrigation.

Yin et al. (2020) applied bi-directional LSTM (Bi-LSTM) to forecast short term reference ET (one day lead time) in areas with limited meteorological data by using inputs of maximum, minimum temperature, sunshine duration and observed that sunshine duration has a higher correlation with reference ET than solar radiation. Hence including the sunshine duration in the forecast model can improve model accuracy. This study was also able to the ability of Bi-LSTM to represents the temporal variability of reference ET over the year.

For our study, in terms of accuracy, the forecast showed a greater skill for irrigated crops compared to non-irrigated crops. We also found higher accuracy for coarse-grained soils (sandy loam, loam). Results suggest that developed forecasting models are promising for simulating ET in the growing season, but the methods need to be improved for fine-textured and non-irrigated conditions. The performance of ET forecasting can be improved by selecting appropriate meteorological parameters as the input features of the model. At the same time, ET had strong regional characteristics such as different accuracy for different soil types. Future work will involve testing how such forecasts could be directly implemented for irrigation management and what changes can be made to reduce model bias.

3.4.4 Limitations and future directions

ML models such as RF and LSTM models show better generalization than linear models and can perform well in space and time compared to one-layer ANNs or autoregressive models (Fang et al., 2017). While ML models are useful for ET modeling, they have limits. For example, the models here are locally calibrated. While the calibration was pooled across multiple crop and soil types, it is possible that some combinations of crops and soils were not well trained and could lead to inaccurate prediction of ET at those locations. Significant training data is a limitation to the ML models. Long-term climatic data can help data-driven models to extract the climatic cycle influence on ET. Hence models developed on those domains with long-term flux tower locations would be more reliable to predict ET and less sensitive to uncertainty than those regions with shorter-term and fewer ET data. In cases with limited training data, mechanistic models do have an indisputable advantage of estimating hydrological variables for any set of inputs as long as the limitations and assumptions of the model are valid.

In terms of parameters, one limitation of our proposed model is the lack of root zone water dynamics. For example, when soils have enough water stored in them during the wet year, actual ET under non-irrigated conditions is assumed to be equal to the potential crop ET. However, during dry conditions, limited soil water storage is often observed, which can reduce actual ET, and plant ET is more a function of soil moisture. We also observed that soil moisture proxy predictors (in form of prcp 7 and prcp 30) were of particular importance for non-irrigated crops. This could be because spells of heat waves during dry years (e.g., 2012, 2010) can lead to a more rapid decline in soil moisture in non-irrigated sites compared to irrigated sites. ML models

also tend to perform poorly on extrapolation to conditions not observed in the data or during extreme or rare events. We saw these results in Fig. 3.14 for extreme events in a dry year (2012) and a wet year (2017). The tendency of all ML models to “regress to the mean” limits their usefulness in flash drought or flooding type conditions that may become more prevalent with ongoing anthropogenic climate change. In addition, Gupta et al, (2009) also found that this result is more expected when using MSE as a calibration objective function.

The future application of LSTM and RF models will be catalyzed with the availability of more data under more conditions. There is also promising research in improving the representation of processes within ML, using reinforcement learning or physical constraint type approaches (Zhao et al., 2019). For example, it is possible to add physical properties to account for subsurface dynamics by including an additional input layer of tree nodes. Even though the proposed model does not have a representation of water balance, it is possible to link neurons and trees to atmospheric and hydrological patterns, such as heat fluxes, so that water is conserved and allowing for less realistic ET estimates to be rejected. However, this might come at a cost of requiring more input predictors that must be derived from data products that may or may not be available. It is also possible to physically constraint ML models (Gorman and Dwyer et al., 2018, Zaherpour et al., 2019 and Zhao et al., 2019b), which can help to conserve energy budget while accounting for physical transport processes of water vapor, leading to a better generalization of physical processes during extremes. Camporeale (2019) also underscores the need to do more research into probabilistic-based uncertainty estimates and the development of gray box models by combining mechanistic and ML approaches.

It will also be useful to collect more data from other climate regimes, crops, and soil types that can help us understand if the conclusions found here and related papers can be generalized to other regions and other crops. This can be used to study the scale- and location- dependence of the drivers on ET and help improve ET forecasting in regional scales.

3.5 Conclusion

We proposed a new framework based on a machine learning data-driven network to estimate and forecast ET and its uncertainty for corn, soybeans, and potatoes under different soil texture types in agricultural areas of the Midwest, USA. The model was built by using biophysical and meteorological information acquired from ground observations and satellite sensor data. The data sets used in the proposed model have been widely utilized in many studies for ET prediction and related to ancillary data used in hydrological models such as SWAT and HSPF. The proposed model was calibrated using 13 field-based eddy covariance ET time series distributed across the region for the period of 2003-2019 for irrigated and rainfed agricultural areas in the Midwest. The model was evaluated in seven independent locations for the time period of 2003-2019.

The evaluation results based on observed ET measurements collected from seven different sites confirmed that the predicted models can be used for daily ET estimates with ubRMSE from 0.67-0.92 mm, Willmott's skill score from 0.80-0.90 and simulate the spatial heterogeneity of agricultural parameters and dynamics of water use by crops. The prediction model estimates were reliable and on par with mechanistic model estimates from NLDAS. The results of this study also revealed that the inclusion of EVI, solar zenith angle, incoming SW

radiation, and CumGDD were the most important input predictors. Vapor pressure was of greater importance for forecasting future ET. The proposed model can also be applied to both rainfed and irrigated crop types. Overall, our work supports the use of ML, especially random forest approaches for prediction and short-term forecasting of ET in both rainfed and irrigated crops, which had a range of valuable uses for irrigation management and water cycling evaluation. Expanding this work outward to tropical or semi-arid regions may require further evaluation of additional predictors, but overall, the results here find that a general field-scale regional ET model is realizable across a range of soil characteristics and climatic patterns. ET prediction and forecasting by using this modeling framework can help policy makers to allocate water sustainability for irrigation and assist growers to spot water stress areas in farms.

3.6 Acknowledgements

ARD and AT acknowledge support from the Wisconsin Vegetable and Potato Growers Association award to UW-Madison, the Wisconsin Department of Natural Resources, and the UW Center for Climatic Research Climate, People, and Environment Program, and thank J Thom, T Houlihan, J Pavelski for site support at US-CS1 and US-CS3. DR and JC are supported by the NASA Carbon Cycle & Ecosystems program (NNX17AE16G).

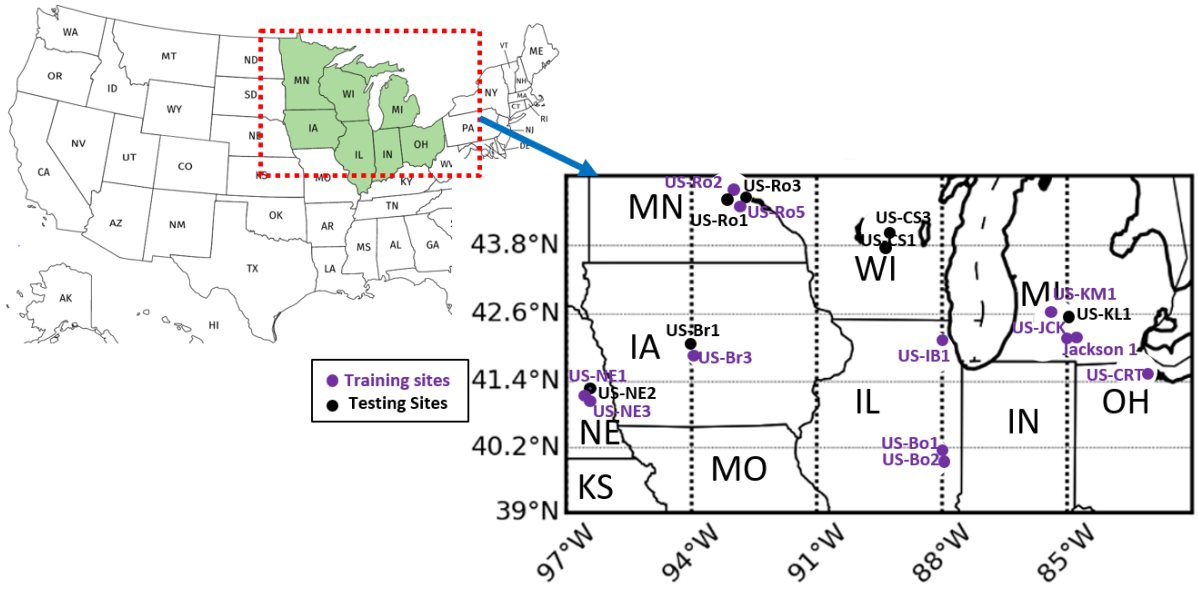

Figure 3.1: Study sites and locations for calibration and evaluation data. AmeriFlux site ID were used to identify locations. Thirteen sites were used for training with some part of data (2009-2016) for calibration and remaining data (2017-2018) from the same sites for testing. Hence thirteen sites (n=18481) were used for training and seven independent sites (n=7850) were used in testing model performance.

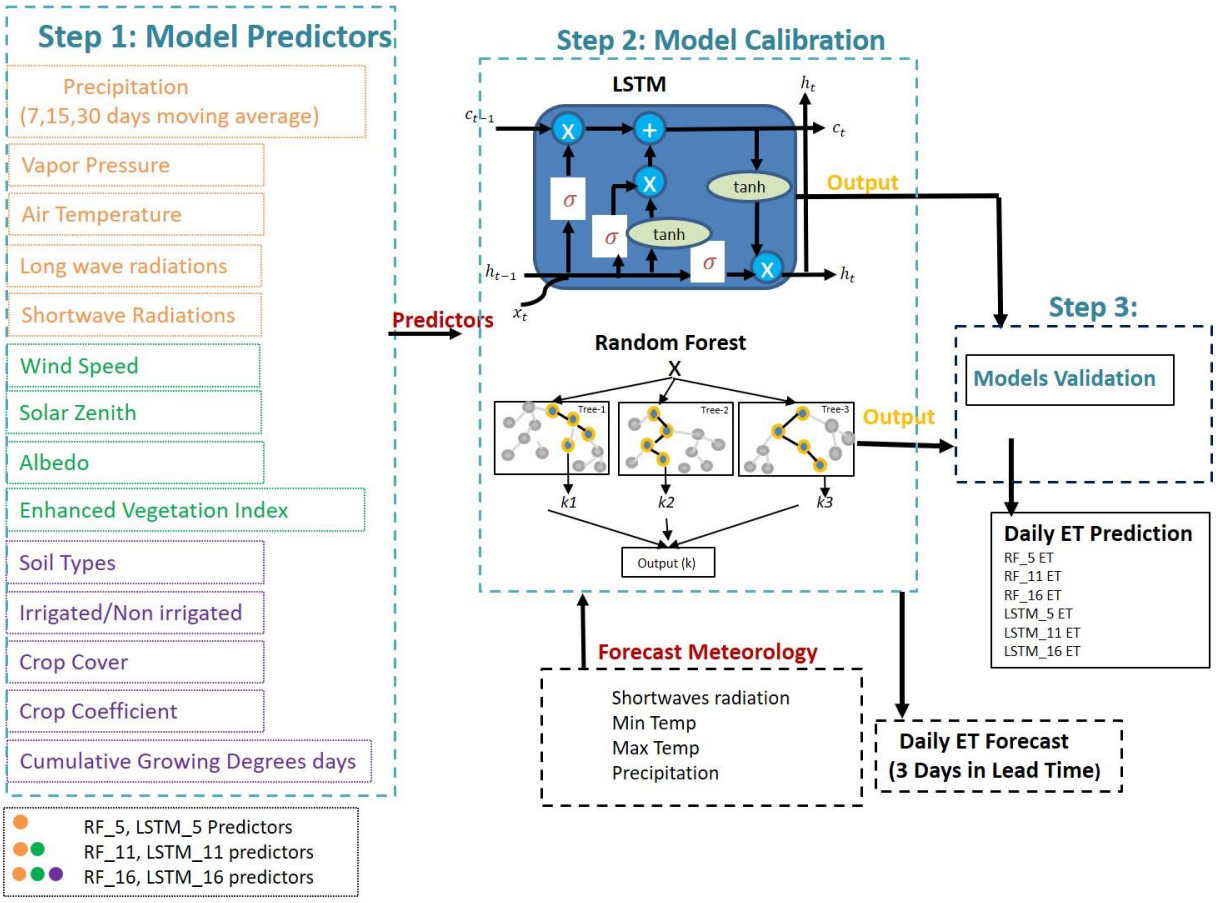

Figure 3.2: Framework of key steps for proposed daily ET prediction and forecast models. RF_5, LSTM_5 model predictors are in orange color, RF_11, LSTM_11 model predictors are in orange and green color, RF_16, LSTM_16 predictors are in orange, green and purple color.

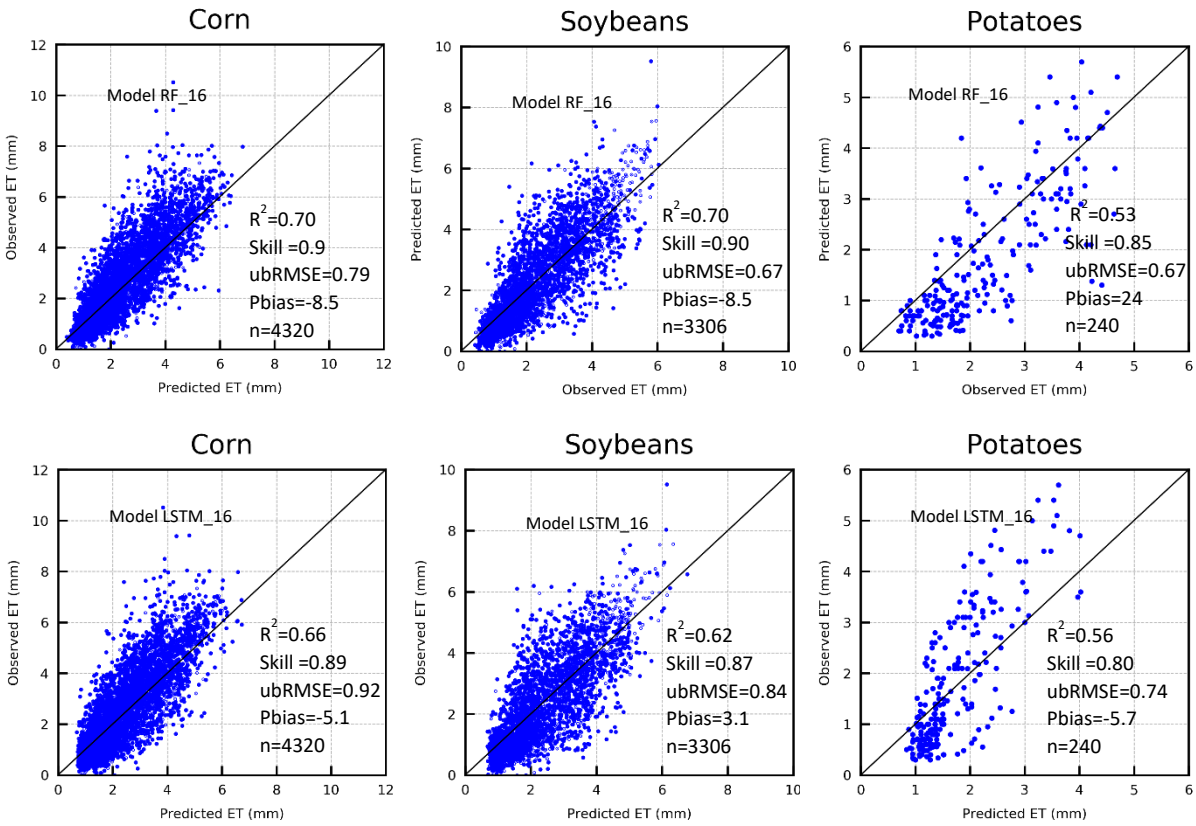

Figure 3.3: Evaluation results of the proposed daily Random Forest (RF_16) and LSTM (LSTM_16) prediction models for various soil texture for flux tower locations in Midwest. 16 means model with 16 variables. n= sample size.

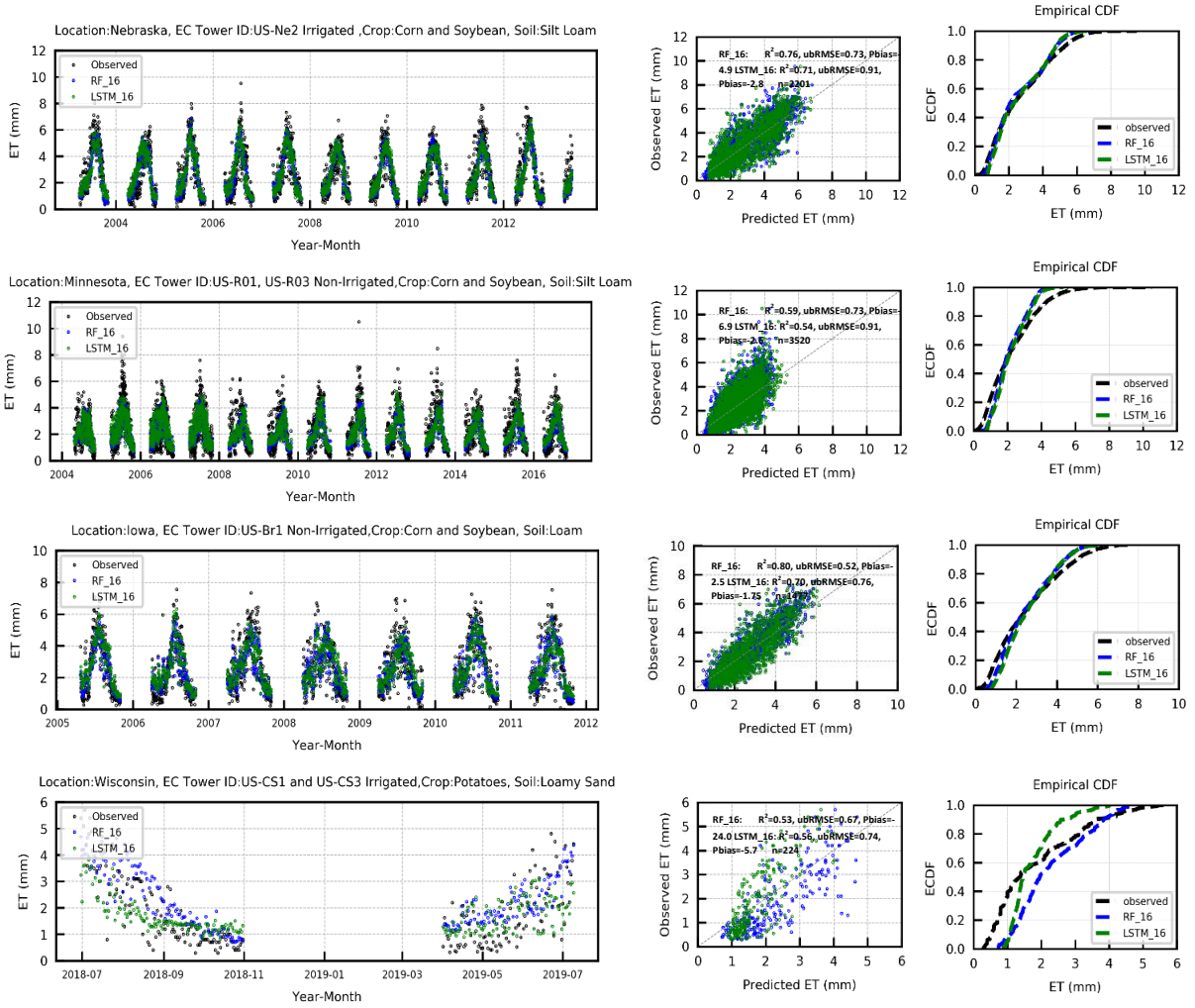

Figure 3.4: Time series, scatter plot, and ECDF of modeled and observed ET; Observed = ET from six representative study sites; Data for US-CS1 and USCS3 were presented together in one graph. The black dots and dotted line show observed ET and the blue and green dotted line and points purple are indicating, respectively RF_16 and LSTM_16 model. RF_16=random forest model with 16 input variables; LSTM_16=LSTM model with 16 input variables; ECDF = Empirical Distribution Function; R^2= Coefficient of determination; RMSE = Root Mean Square Error; Pbias=Percentage bias.

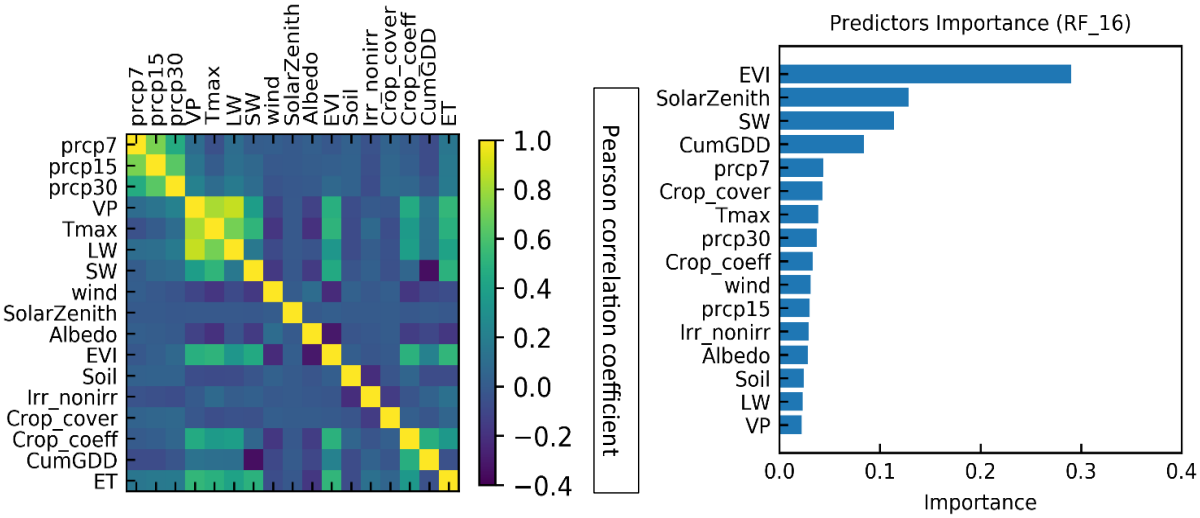

Figure 3.5: Pearson Correlation coefficient between prediction model predictors and response (daily ET); Predictors importance for RF_16 model. Values were scaled to 1 to calculate the influence of each predictor on the response.

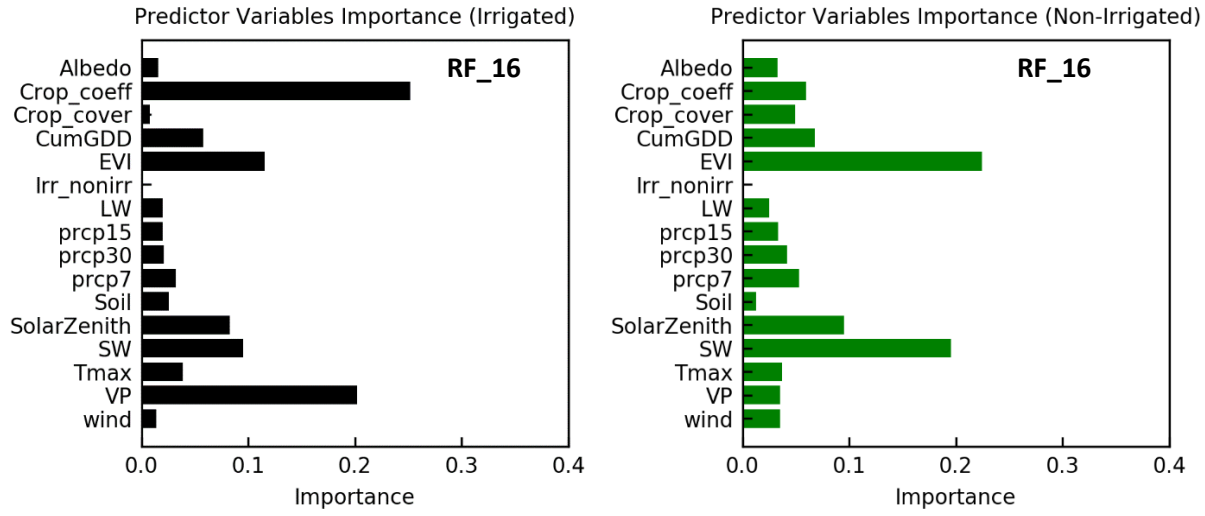

Figure 3.6: Predictors importance for irrigated and non-irrigated crops based on RF_16 model. Predictor importance is scaled to one which means that sum of the contribution of all predictors is equal to one. Predictors with longer horizontal bars are more important in terms of explaining model variance.

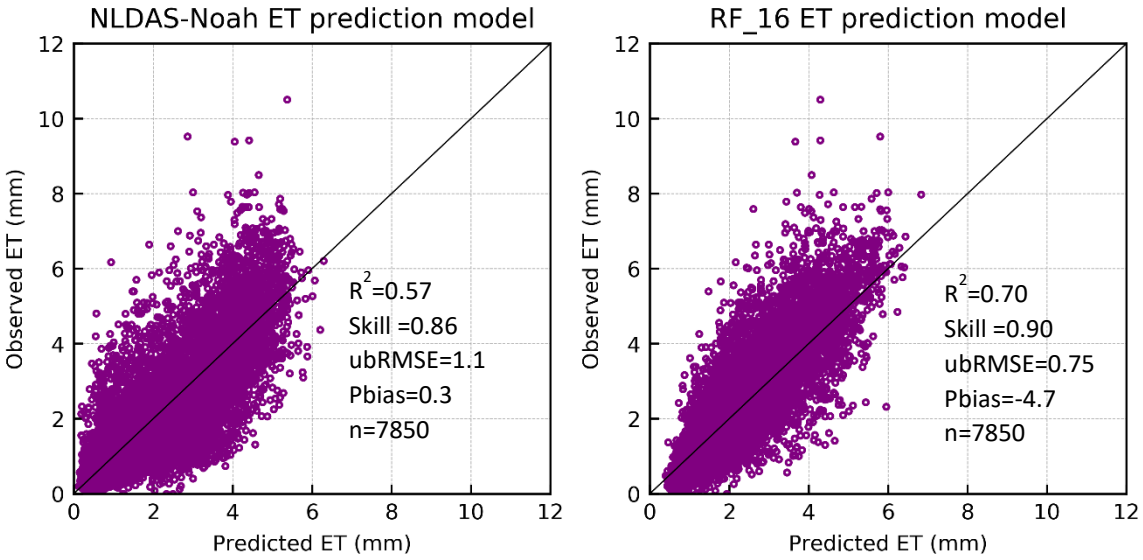

Figure 3.7: Scatter plot of evaluation results of RF and NLDAS-Noah prediction models for sample size n=7850 (30% of whole data). NLDAS-Noah is considered as a mechanistic benchmark model that is compared with the overall best model RF_16.

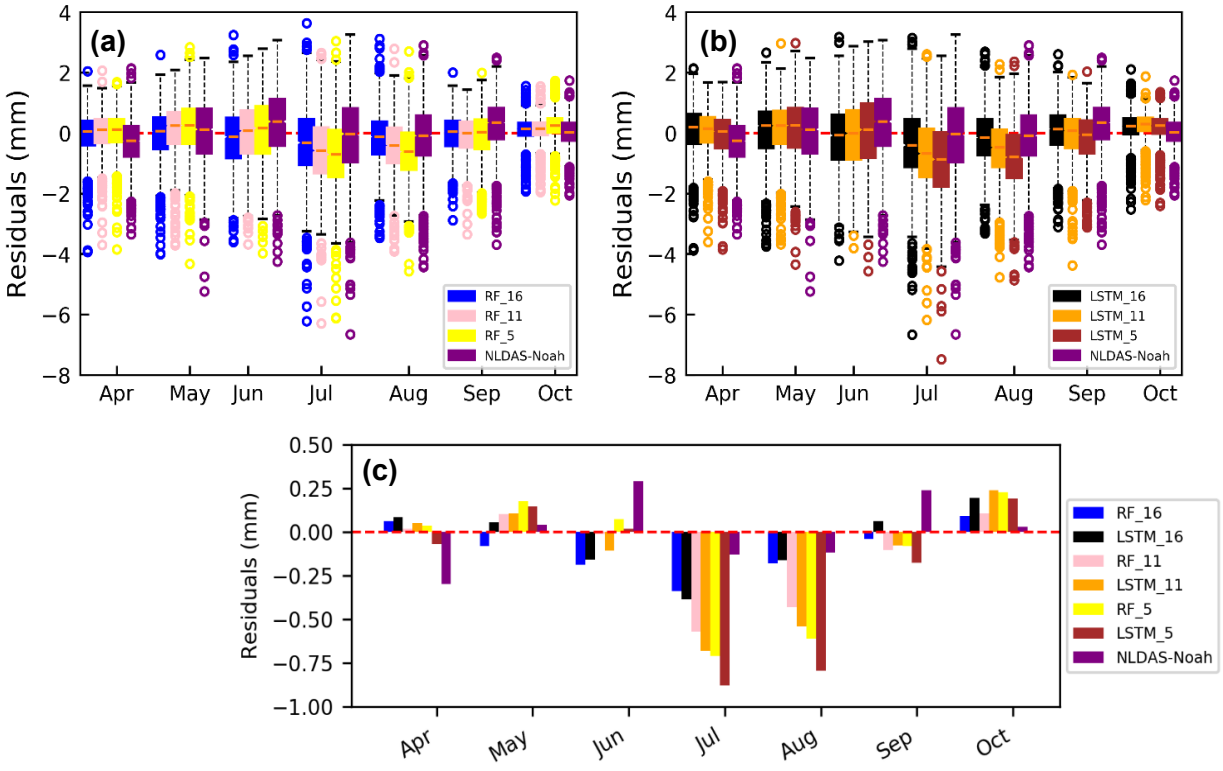

Figure 3.8: Box plots for predicted ET residuals (simulated minus observed) for evaluation results of RF (a) and LSTM (b) prediction models for sample size n=7850 (30% of whole data). RF and LSTM models are also compared with NLDAS-Noah. Median and the 25th and 75th percentiles are represented by boxes. The whiskers represent one and a half times the interquartile range (or ~ ±2.7σ). Circles show outliers. Figure (c) also shows the mean residual comparison between different versions of RF and LSTM models and NLDAS-Noah.

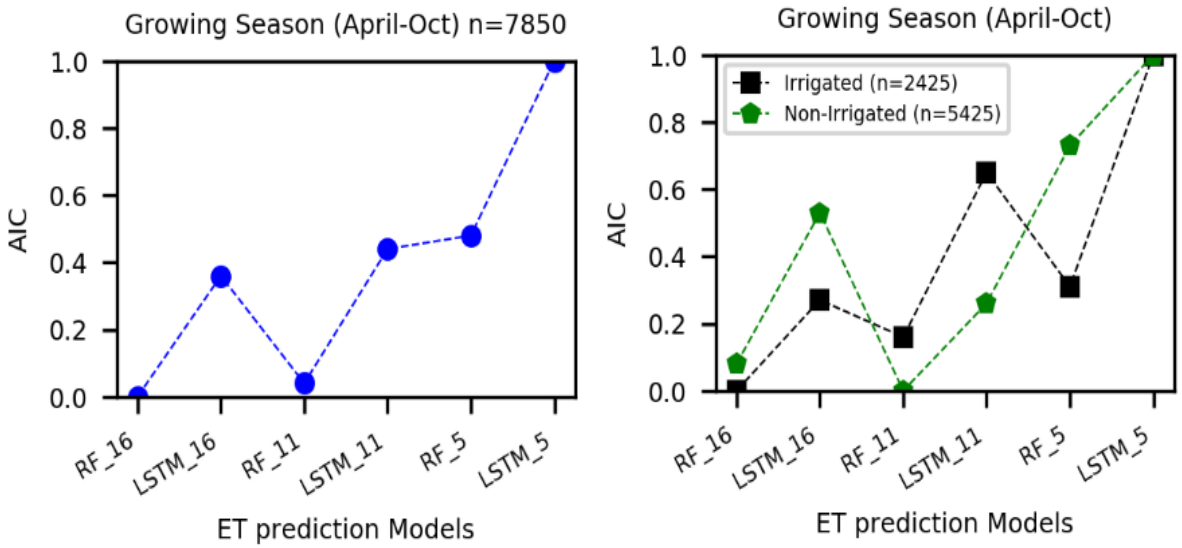

Figure 3.9: AIC scores for different versions of prediction models on evaluation data. n represents the sample size. AIC score is normalized between 0-1 for comparison. First AIC was calculated for the whole data set (n=7850) for different versions of prediction models. Then data are divided into irrigated and non-irrigated crops and AIC is calculated separately for irrigated and non-irrigated crops because of different sample sizes.

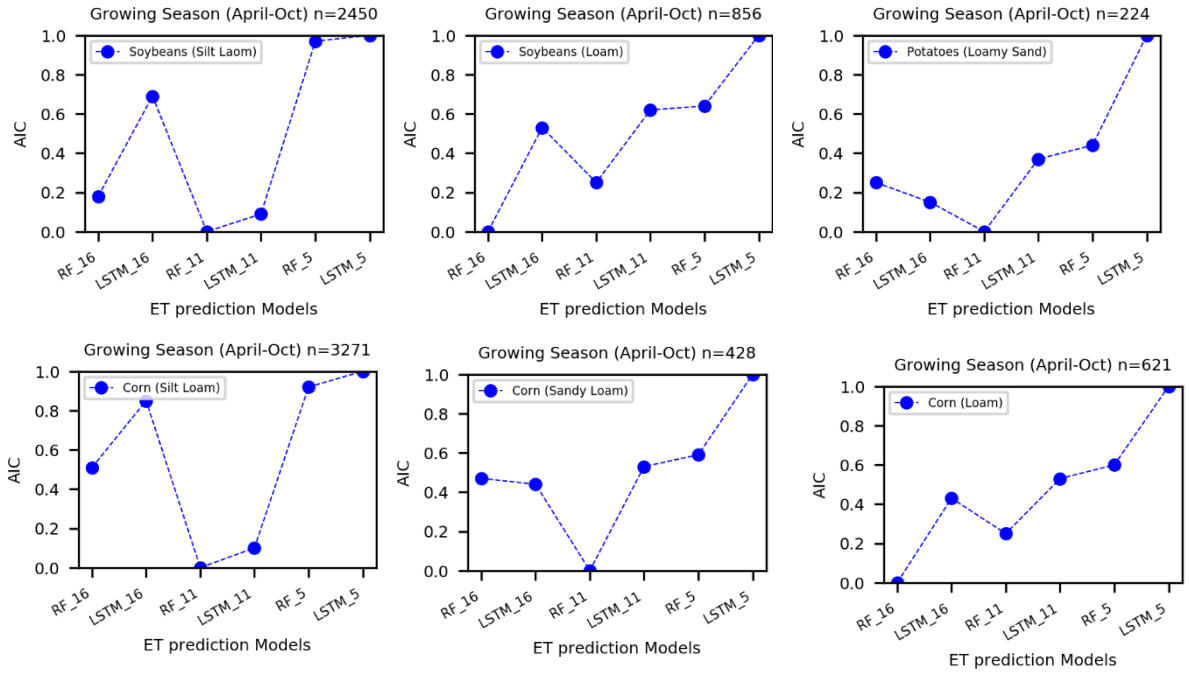

Figure 3.10: AIC scores for different versions of prediction models on evaluation data based on soil types and crop types. n represents the sample size. AIC score is normalized between 0-1 for comparison to account for different parameter numbers among the models. Lower AIC means the model is more parsimonious than a model with higher AIC.

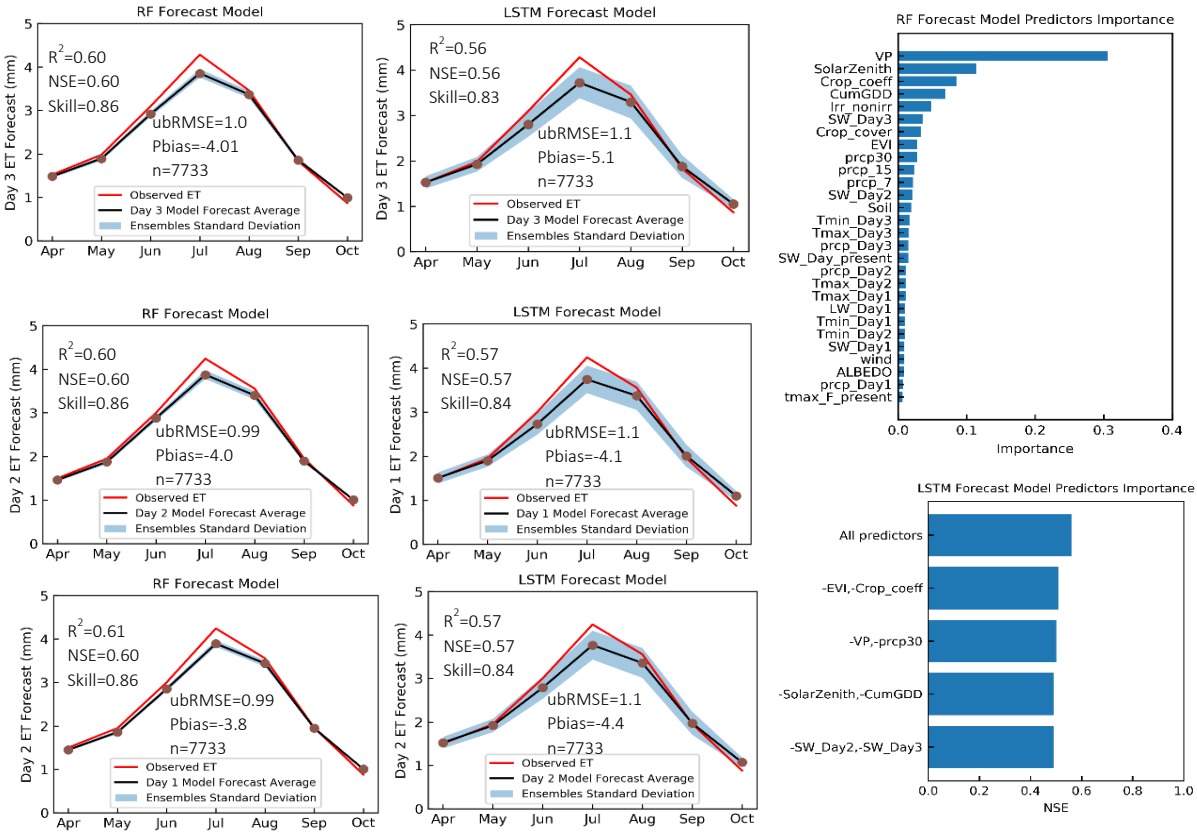

Figure 3.11: RF and LSTM comparison for day1,2,3 ET forecast on evaluation data. ET values are averaged by month for visualization of ensemble spread for each month. Evaluation statistics are calculated on daily ET forecast estimate. Graphs on the right column show RF models and the graph in the middle column shows LSTM models. Variable importance is shown for day 3 ET forecast in the graph on the right column for RF and LSTM. The effect of only important variable removal on NSE was shown for the LSTM model, while the importance of all predictors is shown for the RF forecast model.

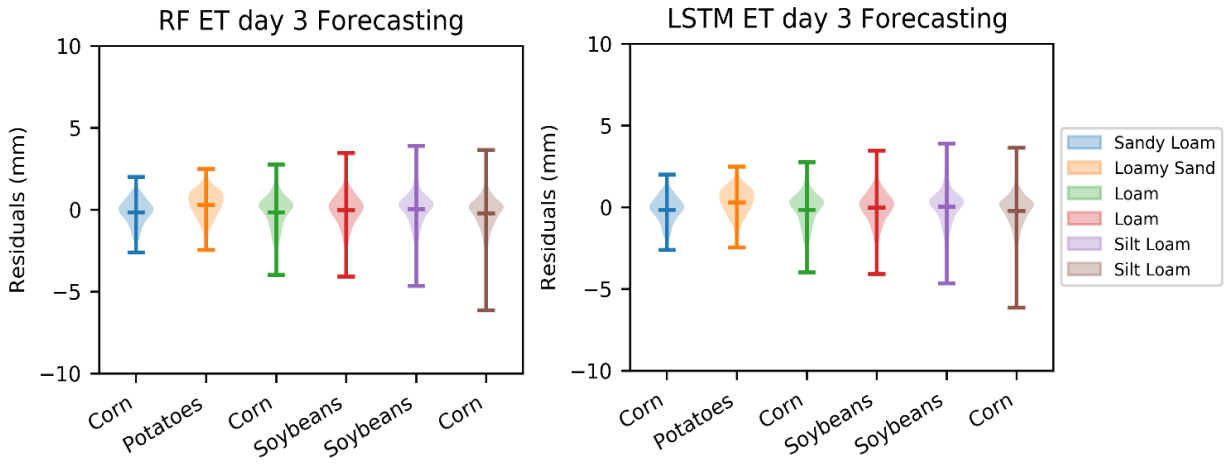

Figure 3.12: Day 3 ET forecast for corn, soybeans, and potatoes under different soil textures for evaluation RF and LSTM. Different colors show the combination of crop types with soil textures.

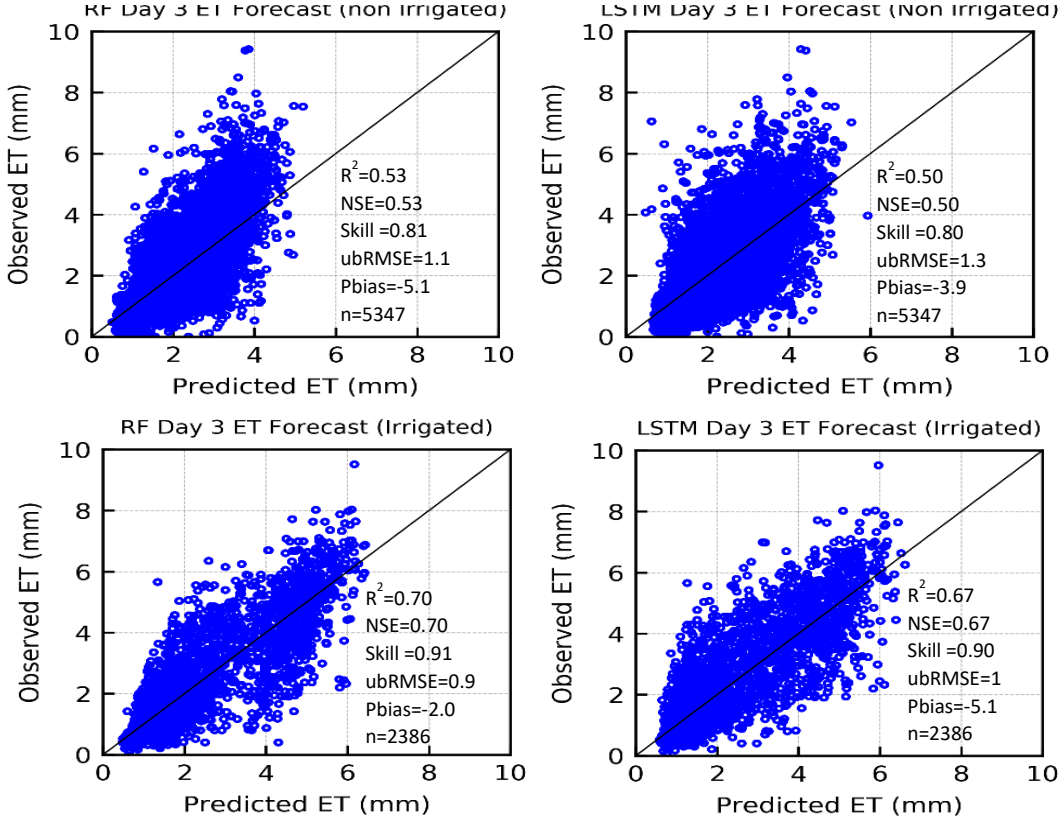

Figure 3.13: Comparison of R^2, MAE and Pbias for day 3 ET forecast results. Both RF and LSTM models were tested in irrigated and non-irrigation crops. n represents the sample size of evaluation data.

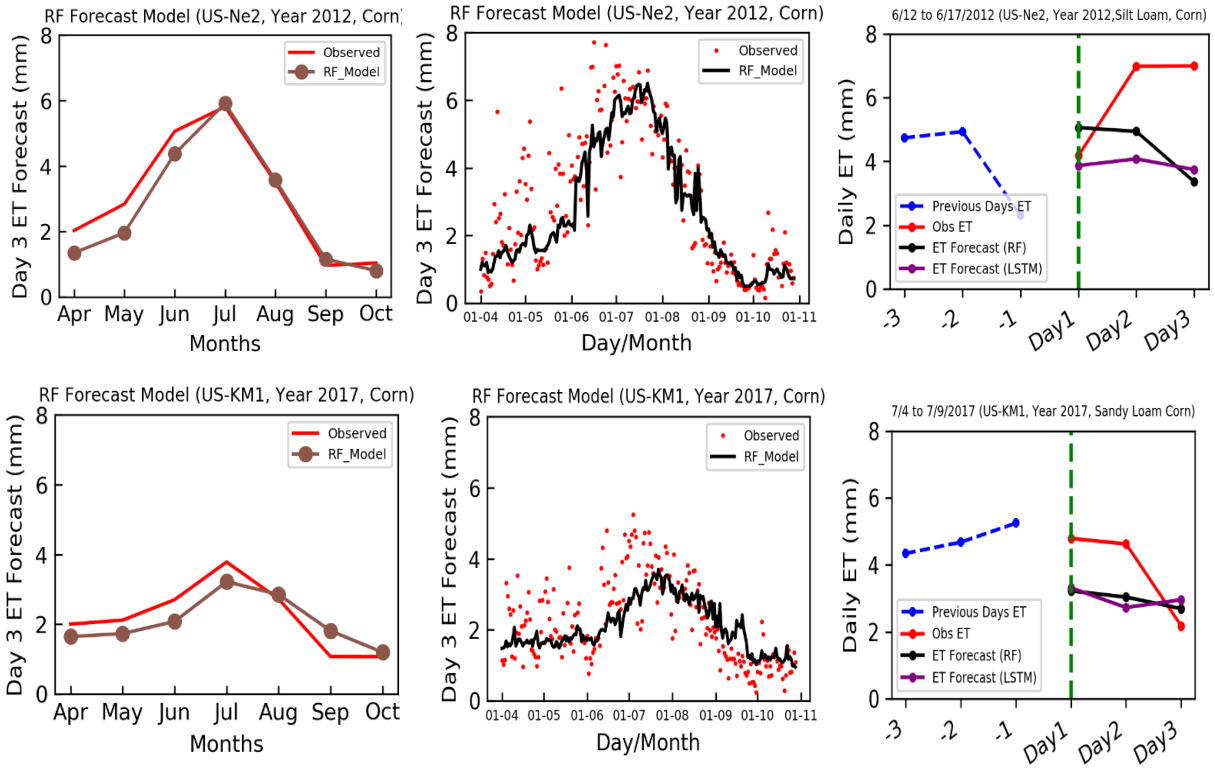

Figure 3.14: RF and LSTM model Day 3 ET forecast performance on evaluation data for a dry year (2012) with flash drought and wet year (2017). Graphs on the top and bottom right show the performance of RF and LSTM models for forecasting day1, day 2, and day 3 ET in 2012 (US-Ne2) and 2017 (US-KM1). Graphs in the middle show comparison based on daily data for growing season and graphs on left shows average monthly ET.

Table 3.1*:* Description of agricultural flux towers study sites located in Midwest USA. Sites names are based on AmeriFlux

State	Site ID	Lat.	Long.	Duration	Soil Type	Rain-Fed/Irrigated	Crop types	Doi
IA	US-Br1	41.97	-93.69	2005-2011	Loam	Rain-Fed	Corn in odd years, Soy in even years and 2011	Prueger and Parkin (2001a) https://doi.org/10.17190/AMF/1246038
IA	US-Br3	41.97	-93.69	2005-2011	Clay Loam	Rain-Fed	Corn in even years, Soy in odd years	Prueger and Parkin (2001b) https://doi.org/10.17190/AMF/1246039
IL	US-IB1	41.86	-88.22	2006-2017	Silt Loam	Rain-Fed	Corn in even years and 2013,2017, Soy in odd years and 2014	Matamala, (2005) https://doi.org/10.17190/AMF/1246065
IL	US-Bo2	40.01	-88.29	2004-2007	Silty Clay	Rain-Fed	Corn in even years, Soy in odd years	Bernacchi, (2004-2008) https://doi.org/10.17190/AMF/1246037
IL	US-Bo1	40.01	-88.29	2005-2016 except 2008-2009	Silt Loam	Rain-Fed	Corn in odd years, Soy in even years	Meyers,(1996-) https://doi.org/10.17190/AMF/1246036
MI	US-KL1	42.48	-85.44	2009-2018	Sandy Loam	Rain-Fed	Soy in 2009, Corn in 2010-2018	Chen,J., (2009-2018) https://ameriflux.lbl.gov/sites
MI	US-JCK	42.21	-84.85	2018	Sandy Loam	Rain-Fed	Soy	Chen,J., (2018a) https://ameriflux.lbl.gov/sites
MI	US-KM1	42.44	-85.33	2014-2018	Sandy Loam	Rain-Fed	Corn	Chen,J., (2018b)
MI	Jackson 1	42.26	-84.84	2018	Loamy Sand	Irrigated	Corn	Chen,J., (2018c)
MN	US-Ro1	44.71	-93.09	2004-2016	Silt Loam	Rain-Fed	Corn in odd years, Soy in even years	Baker and Griffis (2003-2017a) https://doi.org/10.17190/AMF/1246092
MN	US-Ro2	44.73	-93.09	2008, 2011,2012, 2016	Silt Loam	Rain-Fed	Soy in 2012, Corn in 2008, 2011,2016	Baker and Griffis (2003-2017b) https://doi.org/10.17190/AMF/1418683
MN	US-Ro3	44.72	-93.09	2004-2007	Silt Loam	Rain-Fed	Corn in odd years, Soy in even years	Baker and Griffis (2003-2013) https://doi.org/10.17190/AMF/1246093
MN	US-Ro5	44.69	-93.06	2017-2018	Silt Loam	Rain-Fed	Soy in 2017, Corn in 2018	Baker and Griffis (2017) https://doi.org/10.17190/AMF/1419508
NE	US-Ne2	41.16	-96.47	2003-2013	Silt Loam	Irrigated	Soy in 2004,2006 and 2008. Corn in other years	Suyker, (2001a) https://doi.org/10.17190/AMF/1246085
NE	US-Ne3	41.18	-96.44	2003-2013	Silt Loam	Rain-Fed	Corn in odd years, Soy in even years	Suyker, (2001b) https://doi.org/10.17190/AMF/1246086
NE	US-Ne1	41.17	-96.48	2003-2012	Silty Clay Loam	Irrigated	Corn	Suyker, (2001c) https://doi.org/10.17190/AMF/1246084
OH	US-CRT	41.63	-83.35	2011-2012	Silt Loam	Rain-Fed	Soy	Chen, (2011-2013) https://doi.org/10.17190/AMF/1246156

WI	US-CS1	44.10	-89.54	2018-2019	Loamy Sand	Irrigated	Potatoes	Desai, (2018-2019) https://doi.org/10.17190/AMF/1617710
WI	US-CS3	44.14	-89.57	2019	Loamy Sand	Irrigated	Potatoes	Desai, (2018-2019) https://doi.org/10.17190/AMF/1617710

Table 3.2*:* Descriptive statistics for agricultural flux towers study sites located in Midwest USA. Mean, standard deviation, sample variance, skewness, minimum and maximum daily ET was calculated to show ET variability across sites.

State	Site ID	Mean	Standard Deviation	Sample Variance	Skewness	Minimum daily ET	Maximum daily ET	Rain-Fed/Irrigated
IA	US-Br1	2.56	1.64	2.70	0.64	0.00	7.56	Rain-Fed
IA	US-Br3	2.50	1.57	2.46	0.63	0.00	9.27	Rain-Fed
IL	US-IB1	2.54	1.38	1.91	0.63	0.19	8.12	Rain-Fed
IL	US-Bo2	1.96	1.61	2.60	0.64	0.00	8.23	Rain-Fed
IL	US-Bo1	1.60	1.26	1.60	0.93	0.00	6.81	Rain-Fed
MI	US-KL1	2.22	1.13	1.27	0.81	0.11	6.86	Rain-Fed
MI	US-JCK	1.75	0.77	0.60	0.45	0.30	3.85	Rain-Fed
MI	US-KM1	2.15	1.13	1.28	0.81	0.14	7.04	Rain-Fed
MI	Jackson 1	2.10	1.30	1.68	1.93	0.00	7.99	Irrigated
MN	US-Ro1	2.32	1.51	2.29	0.88	0.00	10.51	Rain-Fed
MN	US-Ro2	2.55	1.25	1.57	0.27	0.08	6.97	Rain-Fed
MN	US-Ro3	2.13	1.30	1.69	0.86	0.00	7.19	Rain-Fed
MN	US-Ro5	2.06	1.24	1.54	0.79	0.28	6.64	Rain-Fed
NE	US-Ne2	2.75	1.77	3.13	0.60	0.08	9.52	Irrigated
NE	US-Ne3	2.46	1.62	2.62	0.57	0.09	6.85	Rain-Fed
NE	US-Ne1	2.86	1.85	3.42	0.53	0.00	9.77	Irrigated
OH	US-CRT	2.83	1.60	2.55	0.93	0.51	8.87	Rain-Fed
WI	US-CS1	1.82	1.29	1.66	0.90	0.30	5.70	Irrigated
WI	US-CS3	1.70	1.25	1.63	0.92	0.28	5.30	Irrigated

Table 3.3: Description of model inputs and predictors used for different versions of models are included. Number at end of each model name shows the number of predictors used to build model. e.g RF_16 is RF model with 16 predictors and LSTM_5 is LTM model with five predictors. Data sources are included.

Model	Driver	Abbreviation	Source
RF_5 RF_11 RF_16 LSTM_5 LSTM_11 LSTM_16	Precipitation	Prcp7, Prcp15, Prcp30	Daymet
	Vapor pressure	VP	Daymet
	Air Temperature	Tmax, Tavg	Daymet
	Long wave radiation	LW	NLDAS_Forcing LSM
	Shortwave radiation	SW	NLDAS_Forcing LSM
RF_11 RF_16 LSTM_11 LSTM_16	Wind Speed	Wind	Rain-Fed
	Solar Zenith	SolarZenith	Aqua MODIS MYD09GA
	Albedo	Albedo	Aqua MODIS MYDTBGA
	Enhanced vegetation Index	EVI	Aqua MODIS MYD09GA
RF_16 LSTM_16	Soil types	Soil	Soil Survey Geographic Database (SSURGO)
	Irrigated-non irrigated	Irr_nonirr	Ameri flux sites
	Crop Cover	Crop_cover	Ameri flux sites
	Crop Coefficient	Crop coeff	Computed as function of growing degrees days
	Cumulative growing degrees days	CumGDD	Computed by empirical formula based on temperature

Table 3.4: Parameters for different version of RF and LSTM prediction models. For RF version, parameters of n_estimators, min_samples_leaf included. LSTM model versions are calibrated using layer 1 dropout, layer 2 neurons, and epoch.

Model	n_estimator	min_samples_leaf	min_samples_split=8
RF_5	100	5	8
RF_11	100	5	8
RF_16	150	5	6
Model	Layer 1 dropout	Layer 2 Neurons	Epoch
LSTM_5	0.25	25	100
LSTM_11	0.25	50	100
LSTM_16	0.1	100	65

Table 3.5: Model performance evaluation statistics for different versions of prediction models on daily timestamp. The number at end of each model name shows the number of predictors used to build model. e.g RF_16 is RF model with 16 predictors and LSTM_5 is LTM model with five predictors.

Model Evaluation Statistics for prediction model	R^2	NSE	Willmott skill score	Pearson Corr.	MAE	ubRMSE (mm/day)	RSR (mm/day)	Pbias (%)	AIC
RF_16	0.70	0.70	0.90	0.84	0.64	0.75	0.55	-4.7	0.0
LSTM_16	0.65	0.65	0.88	0.81	0.72	0.89	0.59	-1.9	0.34
NLDAS_Noah	0.57	0.57	0.86	0.76	0.79	1.1	0.65	0.3	Benchmark
RF_11	0.70	0.70	0.89	0.85	0.66	0.76	0.55	-5.3	0.04
LSTM_11	0.63	0.63	0.86	0.82	0.73	0.91	0.60	-6.0	0.42
RF_5	0.63	0.63	0.85	0.81	0.73	0.94	0.61	-5.4	0.46
LSTM_5	0.53	0.53	0.80	0.75	0.82	1.20	0.69	-9.3	0.94

Table 3.6: Daily ET prediction Models (RF_16, LSTM_16, RF_11, LSTM_11, RF_5, LSTM_5) performance for different soil types and crops based on R^2, ubRMSE and AIC. n=sample size

Prediction Models	Crop Types	Soil Types	Sample Size (n)	R^2	NSE	Willmott's skill score	Pearson Corr.	MAE (mm/day)	ubRMSE (mm/day)	RSR	Pbias (%)	AIC
RF_16	Soybeans	Silt Loam	2450	0.63	0.63	0.87	0.79	0.65	0.76	0.61	-0.61	0.18
LSTM_16	Soybeans	Silt Loam	2450	0.56	0.56	0.85	0.75	0.73	0.89	0.66	3.5	0.69
RF_11	Soybeans	Silt Loam	2450	0.65	0.65	0.88	0.81	0.65	0.71	0.59	1.36	0.00
LSTM_11	Soybeans	Silt Loam	2450	0.61	0.61	0.85	0.79	0.79	0.80	0.62	0.97	0.09
RF_5	Soybeans	Silt Loam	2450	0.56	0.56	0.85	0.75	0.74	0.89	0.66	2.1	0.97
LSTM_5	Soybeans	Silt Loam	2450	0.48	0.48	0.80	0.69	0.79	1.1	0.72	-2.3	1.00
RF_16	Soybeans	Loam	856	0.84	0.84	0.95	0.92	0.51	0.43	0.4	-0.6	0.00
LSTM_16	Soybeans	Loam	856	0.75	0.75	0.91	0.87	0.65	0.68	0.5	2.0	0.53
RF_11	Soybeans	Loam	856	0.80	0.80	0.93	0.92	0.56	0.53	0.45	-4.6	0.25
LSTM_11	Soybeans	Loam	856	0.72	0.72	0.90	0.88	0.67	0.73	0.53	-5.6	0.62
RF_5	Soybeans	Loam	856	0.71	0.71	0.89	0.88	0.67	0.76	0.53	-4.9	0.64
LSTM_5	Soybeans	Loam	856	0.61	0.61	0.85	0.80	0.80	1.0	0.63	-7.8	1.00
RF_16	Potatoes	Loamy Sand	224	0.53	0.53	0.85	0.80	0.72	0.67	0.69	24	0.25
LSTM_16	Potatoes	Loamy Sand	224	0.56	0.56	0.80	0.80	0.69	0.74	0.66	-5.7	0.15
RF_11	Potatoes	Loamy Sand	224	0.58	0.58	0.84	0.79	0.70	0.68	0.65	13.1	0.00
LSTM_11	Potatoes	Loamy Sand	224	0.50	0.50	0.80	0.69	0.74	0.89	0.73	1.83	0.37
RF_5	Potatoes	Loamy Sand	224	0.42	0.42	0.7	0.65	0.77	0.99	0.76	1.45	0.44
LSTM_5	Potatoes	Loamy Sand	224	0.42	0.42	0.75	0.65	0.77	0.99	0.76	1.45	1.00
RF_16	Corn	Silt Loam	3271	0.70	0.70	0.90	0.85	0.68	0.86	0.56	-9.6	0.51
LSTM_16	Corn	Silt Loam	3271	0.66	0.66	0.88	0.82	0.78	1.0	0.59	-6.6	0.85
RF_11	Corn	Silt Loam	3271	0.69	0.69	0.88	0.96	0.70	0.88	0.56	-10.7	0.00
LSTM_11	Corn	Silt Loam	3271	0.62	0.62	0.85	0.83	0.78	1.1	0.62	-12	0.10
RF_5	Corn	Silt Loam	3271	0.63	0.63	0.86	0.84	0.76	1.0	0.61	-10.4	0.92
LSTM_5	Corn	Silt Loam	3271	0.53	0.53	0.8	0.79	0.87	1.3	0.68	-15	1.00
RF_16	Corn	Sandy Loam	428	0.66	0.66	0.88	0.82	0.51	0.45	0.58	-5.1	0.47
LSTM_16	Corn	Sandy Loam	428	0.66	0.66	0.88	0.82	0.53	0.45	0.58	-2.8	0.44
RF_11	Corn	Sandy Loam	428	0.7	0.7	0.9	0.84	0.49	0.41	0.55	2.9	0.00
LSTM_11	Corn	Sandy Loam	428	0.64	0.64	0.88	0.81	0.53	0.47	0.60	5.3	0.53
RF_5	Corn	Sandy Loam	428	0.63	0.63	0.87	0.80	0.54	0.50	0.61	1.6	0.59
LSTM_5	Corn	Sandy Loam	428	0.58	0.58	0.85	0.76	0.59	0.57	0.65	-0.42	1.00
RF_16	Corn	Loam	621	0.76	0.76	0.92	0.88	0.61	0.65	0.49	-5.1	0.00
LSTM_16	Corn	Loam	621	0.68	0.68	0.89	0.83	0.75	0.87	0.56	1.4	0.43
RF_11	Corn	Loam	621	0.71	0.71	0.89	0.88	0.66	0.75	0.54	-9.8	0.25
LSTM_11	Corn	Loam	621	0.65	0.65	0.86	0.84	0.75	0.92	0.59	-7.3	0.53
RF_5	Corn	Loam	621	0.63	0.63	0.85	0.85	0.75	0.97	0.61	-10.3	0.60
LSTM_5	Corn	Loam	621	0.52	0.52	0.79	0.77	0.87	1.3	0.69	-11	1.00

Table 3.7: Comparison of RF and LSTM model for day1, day2 and day 3 ET forecast. Models are evaluated on 7733 daily ET observations (2003-2019) from seven sites in Midwest. Model evaluation statistics are calculated on daily timestep.

Model Evaluation Statistics for prediction model	R^2	NSE	Willmott's skill score	Pearson Corr.	MAE	ubRMSE (mm/day)	RSR	pbias (%)
RF_day 1	0.61	0.61	0.86	0.78	0.75	0.99	0.63	-3.8
RF_day 2	0.60	0.60	0.86	0.78	0.75	0.99	0.63	-4.0
RF_day 3	0.60	0.60	0.86	0.78	0.75	1.01	0.63	-4.1
LSTM_day 1	0.57	0.57	0.84	0.76	0.79	1.1	0.65	-4.4
LSTM_day 2	0.57	0.57	0.84	0.76	0.79	1.1	0.65	-4.1
LSTM_day 3	0.56	0.56	0.84	0.76	0.80	1.1	0.66	-5.1

3.7 Bibliography

Abdullah, S.S., Malek, M.A., Abdullah, N.S., Kisi, O., Yap, K.S., 2015. Extreme Learning Machines: A new approach for prediction of reference evapotranspiration. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2015.04.073

Adhikari, A., Ehsani, M.R., Song, Y., Behrangi, A., 2020. Comparative Assessment of Snowfall Retrieval From Microwave Humidity Sounders Using Machine Learning Methods. Earth Sp. Sci. https://doi.org/10.1029/2020EA001357

Akaike, H., 1970. Statistical predictor identification. Ann. Inst. Stat. Math. https://doi.org/10.1007/BF02506337

Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop evapotranspiration: Guidelines for computing crop requirements. Irrig. Drain. Pap. No. 56, FAO. https://doi.org/10.1016/j.eja.2010.12.001

Anandhi, A., 2016. Growing degree days - Ecosystem indicator for changing diurnal temperatures and their impact on corn growth stages in Kansas. Ecol. Indic. https://doi.org/10.1016/j.ecolind.2015.08.023

Anderson, M.C., Kustas, W.P., Norman, J.M., Hain, C.R., Mecikalski, J.R., Schultz, L., González-Dugo, M.P., Cammalleri, C., D'Urso, G., Pimstein, A., Gao, F., 2011. Mapping daily evapotranspiration at field to continental scales using geostationary and polar orbiting satellite imagery. Hydrol. Earth Syst. Sci. https://doi.org/10.5194/hess-15-223-2011

Anderson, M.C., Yang, Yang, Xue, J., Knipper, K.R., Yang, Yun, Gao, F., Hain, C.R., Kustas, W.P., Cawse-Nicholson, K., Hulley, G., Fisher, J.B., Alfieri, J.G., Meyers, T.P., Prueger, J., Baldocchi, D.D., Rey-Sanchez, C., 2021. Interoperability of ECOSTRESS and Landsat for mapping evapotranspiration time series at sub-field scales. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2020.112189

Auret, L., Aldrich, C., 2012. Interpretation of nonlinear relationships between process variables by use of random forests. Miner. Eng. https://doi.org/10.1016/j.mineng.2012.05.008

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X., Malhi, Y., Meyers, T., Munger, W., Oechel, W., Paw, U.K.T., Pilegaard, K., Schmid, H.P., Valentini, R., Verma, S., Vesala, T., Wilson, K., Wofsy, S., 2001. FLUXNET: A New Tool to Study the Temporal and Spatial Variability of Ecosystem-Scale Carbon Dioxide, Water Vapor, and Energy Flux

Densities. Bull. Am. Meteorol. Soc. https://doi.org/10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2

Barr, A.G., van der Kamp, G., Black, T.A., McCaughey, J.H., Nesic, Z., 2012. Energy balance closure at the BERMS flux towers in relation to the water balance of the White Gull Creek watershed 1999-2009. Agric. For. Meteorol. https://doi.org/10.1016/j.agrformet.2011.05.017

Barr, C., Tibby, J., Gell, P., Tyler, J., Zawadzki, A., Jacobsen, G.E., 2014. Climate variability in south-eastern Australia over the last 1500 years inferred from the high-resolution diatom records of two crater lakes. Quat. Sci. Rev. https://doi.org/10.1016/j.quascirev.2014.05.001

[dataset] Baker,J., Griffis ,T., 2003-2010. AmeriFlux US-Ro3 Rosemount- G19, Dataset. https://doi.org/10.17190/AMF/1246093

[dataset] Baker,J., Griffis,T., 2003-2017a. AmeriFlux US-Ro1 Rosemount- G21, Dataset. https://doi.org/10.17190/AMF/1246092

[dataset] Baker,J., Griffis,T., 2003-2017b. AmeriFlux US-Ro2 Rosemount- G21, Dataset. https://doi.org/10.17190/AMF/1418683

[dataset] Baker,J., Griffis, T., 2017. AmeriFlux US-Ro5 Rosemount I18_South, Dataset. https://doi.org/10.17190/AMF/1419508

Belgiu, M., Drăgu, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J. Photogramm. Remote Sens. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bengio, Y., Simard, P., Frasconi, P., 1994. Learning Long-Term Dependencies with Gradient Descent is Difficult. IEEE Trans. Neural Networks. https://doi.org/10.1109/72.279181

Bodesheim, P., Jung, M., Gans, F., Mahecha, M.D., Reichstein, M., 2018. Upscaled diurnal cycles of land-Atmosphere fluxes: A new global half-hourly data product. Earth Syst. Sci. Data. https://doi.org/10.5194/essd-10-1327-2018

Boucher, O., Myhre, G., Myhre, A., 2004. Direct human influence of irrigation on atmospheric water vapour and climate. Clim. Dyn. https://doi.org/10.1007/s00382-004-0402-4

Breiman L., 2001. Machine Learning, 45(1), 5–32. Stat. Dep. Univ. California, Berkeley, CA 94720. https://doi.org/10.1023/A:1010933404324

Breiman, L., 2002. Manual on setting up, using, and understanding random forests v3. 1. Tech. Report, http//oz.berkeley.edu/users/breiman, Stat. Dep. Univ. Calif. Berkeley, …. https://doi.org/10.2776/85168

Camporeale, E., 2019. The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting. Sp. Weather. https://doi.org/10.1029/2018SW002061

Carriere, P., Mohaghegh, S., Gaskari, R., 1996. Performance of a Virtual Runoff Hydrograph System. J. Water Resour. Plan. Manag. https://doi.org/10.1061/(ASCE)0733-9496(1996)122:6(421)

[dataset] Chen,J., Chu, H., 2011-2013. AmeriFlux US-CRT Curtice Walter-Berger cropland, Dataset. https://doi.org/10.17190/AMF/1246156

[dataset] Chen,J., 2018c. AmeriFlux Dataset https://ameriflux.lbl.gov/sites

Chaney, N. W., Herman, J. D., Ek, M. B., & Wood, E. F. 2016. Deriving global parameter estimates for the Noah land surface model using FLUXNET and machine learning. Journal of Geophysical Research, *121*(22). https://doi.org/10.1002/2016JD024821

Chen, F., Dudhia, J., 2001. Coupling and advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part I: Model implementation and sensitivity. Mon. Weather Rev. https://doi.org/10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2

Chen, F., Mitchell, K., Schaake, J., Xue, Y., Pan, H.L., Koren, V., Duan, Q.Y., Ek, M., Betts, A., 1996. Modeling of land surface evaporation by four schemes and comparison with FIFE observations. J. Geophys. Res. Atmos. https://doi.org/10.1029/95JD02165

Chen, Z., Zhu, Z., Jiang, H., Sun, S., 2020a. Estimating daily reference evapotranspiration based on limited meteorological data using deep learning and classical machine learning methods. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2020.125286

Chen, Z., Sun, S., Wang, Y., Wang, Q., Zhang, X., 2020b. Temporal convolution-network-based models for modeling maize evapotranspiration under mulched drip irrigation. Comput. Electron. Agric. https://doi.org/10.1016/j.compag.2019.105206

Clark, M. P., Bierkens, M. F. P., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., Pauwels, V. R. N., Cai, X., Wood, A. W., & Peters-Lidard, C. D. 2017. The evolution of process-based hydrologic models: Historical challenges and the collective quest for physical realism. Hydrology and Earth System Sciences, 21(7). https://doi.org/10.5194/hess-21-3427-2017

Cleland, E.E., Chuine, I., Menzel, A., Mooney, H.A., Schwartz, M.D., 2007. Shifting plant phenology in response to global change. Trends Ecol. Evol. https://doi.org/10.1016/j.tree.2007.04.003

Cobaner, M., 2011. Evapotranspiration estimation by two different neuro-fuzzy inference systems. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2010.12.030

Crosbie, R.S., Davies, P., Harrington, N., Lamontagne, S., 2015. Ground truthing groundwater-recharge estimates derived from remotely sensed evapotranspiration: a case in South Australia. Hydrogeol. J. https://doi.org/10.1007/s10040-014-1200-7

[dataset] Desai.A., 2018-2019. AmeriFlux US-CS1 Central Sands Irrigated Agricultural Field, Dataset. https://doi.org/10.17190/AMF/1617710

[dataset] Desai.A., 2019-2020. AmeriFlux US-CS3 Central Sands Irrigated Agricultural Field, Dataset. https://doi.org/10.17190/AMF/1617713

te Beest, D.E., Mes, S.W., Wilting, S.M., Brakenhoff, R.H., van de Wiel, M.A., 2017. Improved high-dimensional prediction with Random Forests by the use of co-data. BMC Bioinformatics. https://doi.org/10.1186/s12859-017-1993-1

Djaman, K., Smeal, D., Koudahe, K., Allen, S., 2020. Hay yield and water use efficiency of alfalfa under different irrigation and fungicide regimes in a semiarid climate. Water (Switzerland). https://doi.org/10.3390/W12061721

Dong, J., Dirmeyer, P.A., Lei, F., Anderson, M.C., Holmes, T.R.H., Hain, C., Crow, W.T., 2020. Soil Evaporation Stress Determines Soil Moisture-Evapotranspiration Coupling Strength in Land Surface Modeling. Geophys. Res. Lett. https://doi.org/10.1029/2020GL090391

Donohue, R.J., McVicar, T.R., Roderick, M.L., 2010. Assessing the ability of potential evaporation formulations to capture the dynamics in evaporative demand within a changing climate. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2010.03.020

E. Vermote, R.W., 2015. MOD09GQ MODIS/Terra Surface Reflectance Daily L2G Global 250m SIN Grid V006 [WWW Document]. Distrib. by NASA EOSDIS L. Process. DAAC. https://doi.org/doi.org/10.5067/MODIS/MOD09GQ.006

Fang, K., Shen, C., Kifer, D., Yang, X., 2017. Prolongation of SMAP to Spatiotemporally Seamless Coverage of Continental U.S. Using a Deep Learning Neural Network. Geophys. Res. Lett. https://doi.org/10.1002/2017GL075619

Fang, W., Huang, S., Huang, Q., Huang, G., Meng, E., Luan, J., 2018. Reference evapotranspiration forecasting based on local meteorological and global climate information screened by partial mutual information. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2018.04.038

FAO. 2015. Chapter 7 - ETc - Dual crop coefficient. Food and Agriculture Organization of the United Nations Retrieved from http://www.fao.org/docrep/x0490e/x0490e0c.

Feng, Y., Cui, N., Gong, D., Zhang, Q., Zhao, L., 2017. Evaluation of random forests and generalized regression neural networks for daily reference evapotranspiration modelling. Agric. Water Manag. https://doi.org/10.1016/j.agwat.2017.08.003

Ferreira, L.B., da Cunha, F.F., de Oliveira, R.A., Fernandes Filho, E.I., 2019. Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM – A new approach. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2019.03.028

Ferreira, L.B., da Cunha, F.F., 2020. Multi-step ahead forecasting of daily reference evapotranspiration using deep learning. Comput. Electron. Agric. https://doi.org/10.1016/j.compag.2020.105728

Filgueiras, R., Almeida, T.S., Mantovani, E.C., Dias, S.H.B., Fernandes-Filho, E.I., da Cunha, F.F., Venancio, L.P., 2020. Soil water content and actual evapotranspiration predictions using regression algorithms and remote sensing data. Agric. Water Manag. https://doi.org/10.1016/j.agwat.2020.106346

Fisher, J.B., Lee, B., Purdy, A.J., Halverson, G.H., Dohlen, et al., 2020. ECOSTRESS: NASA's Next Generation Mission to Measure Evapotranspiration From the International Space Station. Water Resour. Res. https://doi.org/10.1029/2019WR026058

Gupta, H. V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hamill, T.M., Bates, G.T., Whitaker, J.S., Murray, D.R., Fiorino, M., Galarneau, T.J., Zhu, Y., Lapenta, W., 2013. NOAA's second-generation global medium-range ensemble reforecast dataset. Bull. Am. Meteorol. Soc. https://doi.org/10.1175/BAMS-D-12-00014.1

Izadifar, Z., Elshorbagy, A., 2010. Prediction of hourly actual evapotranspiration using neural networks, genetic programming, and statistical models. Hydrol. Process. https://doi.org/10.1002/hyp.7771

Jensen, M.E., Burmann, R.D., Allen, R.G., 1990. Evaporation and irrigation water requirements, ASCE manual and reports on engineering practice.

Jianping, Z., Zhong, Y., Daojie, W., Xinbao, Z., 2002. Climate change and causes in the Yuanmou dry-hot valley of Yunnan, China. J. Arid Environ. https://doi.org/10.1006/jare.2001.0851

Jung, M., Reichstein, M., Margolis, H.A., Cescatti, A., Richardson, A.D., Arain, M.A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law, B.E., Lindroth, A., Merbold, L., Montagnani, L., Moors, E.J., Papale, D., Sottocornola, M., Vaccari, F., Williams, C., 2011. Global patterns of land-atmosphere fluxes of carbon

dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. J. Geophys. Res. Biogeosciences. https://doi.org/10.1029/2010JG001566

Kao, I.F., Zhou, Y., Chang, L.C., Chang, F.J., 2020. Exploring a Long Short-Term Memory based Encoder-Decoder framework for multi-step-ahead flood forecasting. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2020.124631

Kim, D., Lee, W.S., Kim, S.T., Chun, J.A., 2019. Historical Drought Assessment Over the Contiguous United States Using the Generalized Complementary Principle of Evapotranspiration. Water Resour. Res. https://doi.org/10.1029/2019WR024991

Kimball, B.A., Boote, K.J., Hatfield, J.L., Ahuja, L.R., Stockle, C., Archontoulis, S., Baron, C., Basso, B., Bertuzzi, P., Constantin, J., Deryng, D., Dumont, B., Durand, J.L., Ewert, F., Gaiser, T., Gayler, S., Hoffmann, M.P., Jiang, Q., Kim, S.H., Lizaso, J., Moulin, S., Nendel, C., Parker, P., Palosuo, T., Priesack, E., Qi, Z., Srivastava, A., Stella, T., Tao, F., Thorp, K.R., Timlin, D., Twine, T.E., Webber, H., Willaume, M., Williams, K., 2019. Simulation of maize evapotranspiration: An inter-comparison among 29 maize models. Agric. For. Meteorol. https://doi.org/10.1016/j.agrformet.2019.02.037

Kisi, O., Alizamir, M., 2018. Modelling reference evapotranspiration using a new wavelet conjunction heuristic method: Wavelet extreme learning machine vs wavelet neural networks. Agric. For. Meteorol. https://doi.org/10.1016/j.agrformet.2018.08.007

Koster, R.D., Suarez, M.J., 1996. Energy and Water Balance Calculations in the Mosaic LSM, NASA Technical Memorandum 104606, Technical Report Series on Global Modeling and Data Assimilation.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., Nearing, G., 2019. Benchmarking a Catchment-Aware Long Short-Term Memory Network (LSTM) for Large-Scale Hydrological Modeling. Hydrol. Earth Syst. Sci. Discuss. https://doi.org/10.5194/hess-2019-368

Landeras, G., Ortiz-Barredo, A., López, J.J., 2009. Forecasting weekly evapotranspiration with ARIMA and artificial neural network models. J. Irrig. Drain. Eng. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000008

Li, S., Kang, S., Zhang, L., Zhang, J., Du, T., Tong, L., Ding, R., 2016. Evaluation of six potential evapotranspiration models for estimating crop potential and actual evapotranspiration in arid regions. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2016.10.022

Lian, X., Piao, S., Huntingford, C., Li, Y., Zeng, Z., Wang, X., Ciais, P., McVicar, T.R., Peng, S., Ottlé, C., Yang, H., Yang, Y., Zhang, Y., Wang, T., 2018. Partitioning global land evapotranspiration using CMIP5 models constrained by observations. Nat. Clim. Chang. https://doi.org/10.1038/s41558-018-0207-9

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News.

Lohar, D., Pal, B., 1995. The effect of irrigation on premonsoon season precipitation over south West Bengal, India. J. Clim. https://doi.org/10.1175/1520-0442(1995)008<2567:TEOIOP>2.0.CO;2

Long, D., Longuevergne, L., Scanlon, B.R., 2014. Uncertainty in evapotranspiration from land surface modeling, remote sensing, and GRACE satellites. Water Resour. Res. https://doi.org/10.1002/2013WR014581

Meng, C.L., Li, Z.L., Zhan, X., Shi, J.C., Liu, C.Y., 2009. Land surface temperature data assimilation and its impact on evapotranspiration estimates from the common land model. Water Resour. Res. https://doi.org/10.1029/2008WR006971

Moriasi, D.N., Gitau, M.W., Pai, N., Daggupati, P., 2015. Hydrologic and water quality models: Performance measures and evaluation criteria. Trans. ASABE. https://doi.org/10.13031/trans.58.10715

Moratiel, R., Bravo, R., Saa, A., Tarquis, A.M., Almorox, J., 2020. Estimation of evapotranspiration by the Food and Agricultural Organization of the United Nations (FAO) Penman-Monteith temperature (PMT) and Hargreaves-Samani (HS) models under temporal and spatial criteria - a case study in Duero basin (Spain). Nat. Hazards Earth Syst. Sci. https://doi.org/10.5194/nhess-20-859-2020

Narasimhan, B., Srinivasan, R., 2005. Development and evaluation of Soil Moisture Deficit Index (SMDI) and Evapotranspiration Deficit Index (ETDI) for agricultural drought monitoring, in: Agricultural and Forest Meteorology. https://doi.org/10.1016/j.agrformet.2005.07.012

O'Gorman, P.A., Dwyer, J.G., 2018. Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of Climate, Climate Change, and Extreme Events. J. Adv. Model. Earth Syst. https://doi.org/10.1029/2018MS001351

Oliveira, B.S., Moraes, E.C., Carrasco-Benavides, M., Bertani, G., Mataveli, G.A.V., 2018. Improved albedo estimates implemented in the METRIC model for modeling energy balance fluxes and evapotranspiration over agricultural and natural areas in the Brazilian Cerrado. Remote Sens. https://doi.org/10.3390/rs10081181

Otkin, J.A., Anderson, M.C., Hain, C., Svoboda, M., Johnson, D., Mueller, R., Tadesse, T., Wardlow, B., Brown, J., 2016. Assessing the evolution of soil moisture and vegetation conditions during the 2012 United States flash drought. Agric. For. Meteorol. https://doi.org/10.1016/j.agrformet.2015.12.065

Pandey, P.K., Nyori, T., Pandey, V., 2017. Estimation of reference evapotranspiration using data driven techniques under limited data conditions. Model. Earth Syst. Environ. https://doi.org/10.1007/s40808-017-0367-z

Pauwels, V.R.N., Verhoest, N.E.C., De Lannoy, G.J.M., Guissard, V., Lucau, C., Defourny, P., 2007. Optimization of a coupled hydrology-crop growth model through the assimilation of observed soil moisture and leaf area index values using an ensemble Kalman filter. Water Resour. Res. https://doi.org/10.1029/2006WR004942

Payero, J.O., Irmak, S., 2013. Daily energy fluxes, evapotranspiration and crop coefficient of soybean. Agric. Water Manag. https://doi.org/10.1016/j.agwat.2013.06.018

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res.

Perera, K.C., Western, A.W., Nawarathna, B., George, B., 2014. Forecasting daily reference evapotranspiration for Australia using numerical weather prediction outputs. Agric. For. Meteorol. https://doi.org/10.1016/j.agrformet.2014.03.014

PRIESTLEY, C.H.B., TAYLOR, R.J., 1972. On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters. Mon. Weather Rev. https://doi.org/10.1175/1520-0493(1972)100<0081:otaosh>2.3.co;2

Richards, L.A., 1931. Capillary conduction of liquids through porous mediums. J. Appl. Phys. https://doi.org/10.1063/1.1745010

Rangapuram, S.S., Seeger, M., Gasthaus, J., Stella, L., Wang, Y., Januschowski, T., 2018. Deep state space models for time series forecasting, in: Advances in Neural Information Processing Systems.

Rosenberry, D.O., Winter, T.C., Buso, D.C., Likens, G.E., 2007. Comparison of 15 evaporation methods applied to a small mountain lake in the northeastern USA. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2007.03.018

Sahoo, S., Russo, T.A., Elliott, J., Foster, I., 2017. Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. Water Resour. Res. https://doi.org/10.1002/2016WR019933

Scott, R.L., Cable, W.L., Huxman, T.E., Nagler, P.L., Hernandez, M., Goodrich, D.C., 2008. Multiyear riparian evapotranspiration and groundwater use for a semiarid watershed. J. Arid Environ. https://doi.org/10.1016/j.jaridenv.2008.01.001

Seneviratne, S.I., Corti, T., Davin, E.L., Hirschi, M., Jaeger, E.B., Lehner, I., Orlowsky, B., Teuling, A.J., 2010. Investigating soil moisture-climate interactions in a changing climate: A review. Earth-Science Rev. https://doi.org/10.1016/j.earscirev.2010.02.004

Sharma, V., Kilic, A., Irmak, S., 2016. Impact of scale/resolution on evapotranspiration from Landsat and MODIS images. Water Resour. Res. https://doi.org/10.1002/2015WR017772

Shiri, J., 2018. Improving the performance of the mass transfer-based reference evapotranspiration estimation approaches through a coupled wavelet-random forest methodology. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2018.04.042

Shugart, H. H.,1998. Terrestrial ecosystems in changing environments. Cambridge University Press.https://doi.org/10.1016/s0304-3800(99)00031-9

[dataset] Suyker,A., 2001a. AmeriFlux US-Ne2 Mead - irrigated maize-soybean rotation site, Dataset. https://doi.org/10.17190/AMF/1246085

[dataset] Suyker,A., 2001b. AmeriFlux US-Ne3 Mead - rainfed maize-soybean rotation site, Dataset. https://doi.org/10.17190/AMF/1246086

[dataset] Suyker,A., 2001c. AmeriFlux US-Ne1 Mead - rainfed maize-soybean rotation site, Dataset. https://doi.org/10.17190/AMF/1246084

Song, Y., Ma, M., 2011. A statistical analysis of the relationship between climatic factors and the normalized difference vegetation index in China. Int. J. Remote Sens. https://doi.org/10.1080/01431161003801336

Tabari, H., Kisi, O., Ezani, A., Hosseinzadeh Talaee, P., 2012. SVM, ANFIS, regression and climate based models for reference evapotranspiration modeling using limited climatic data in a semi-arid highland environment. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2012.04.007

Tavares, L.D., Saldanha, R.R., Vieira, D.A.G., 2015. Extreme learning machine with parallel layer perceptrons. Neurocomputing. https://doi.org/10.1016/j.neucom.2015.04.018

Thornton, P.E., Thornton, M.M., Mayer, B.W., Wilhelmi, N., Wei, Y., Devarakonda, R., Cook, R.B., 2014. Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 2. Data set. [WWW Document]. Oak Ridge Natl. Lab. Distrib. Act. Arch. Center, Oak Ridge, Tennessee, USA. https://doi.org/http://dx.doi.org/10.3334/ORNLDAAC/1219

Tennant, C., Larsen, L., Bellugi, D., Moges, E., Zhang, L., Ma, H., 2020. The Utility of Information Flow in Formulating Discharge Forecast Models: A Case Study From an Arid Snow-Dominated Catchment. Water Resour. Res. https://doi.org/10.1029/2019WR024908

Trenberth, K.E., Smith, L., Qian, T., Dai, A., Fasullo, J., 2007. Estimates of the global water budget and its annual cycle using observational and model Data. J. Hydrometeorol. https://doi.org/10.1175/JHM600.1

Valipour, M., Gholami Sefidkouhi, M.A., Raeini−Sarjaz, M., 2017. Selecting the best model to estimate potential evapotranspiration with respect to climate change and magnitudes of extreme events. Agric. Water Manag. https://doi.org/10.1016/j.agwat.2016.08.025

Velpuri, N.M., Senay, G.B., Singh, R.K., Bohms, S., Verdin, J.P., 2013. A comprehensive evaluation of two MODIS evapotranspiration products over the conterminous United States: Using point and gridded FLUXNET and water balance ET. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2013.07.013

Vinukollu, R.K., Sheffield, J., Wood, E.F., Bosilovich, M.G., 2012. Multimodel analysis of energy and water fluxes: Intercomparisons between operational analyses, a land surface model, and remote sensing. J. Hydrometeorol. https://doi.org/10.1175/2011JHM1372.1

Walls, S., Binns, A.D., Levison, J., MacRitchie, S., 2020. Prediction of actual evapotranspiration by artificial neural network models using data from a Bowen ratio energy balance station. Neural Comput. Appl. https://doi.org/10.1007/s00521-020-04800-2

Wang, K., Dickinson, R.E., 2012. A review of global terrestrial evapotranspiration: Observation, modeling, climatology, and climatic variability. Rev. Geophys. https://doi.org/10.1029/2011RG000373

Wei, Z., Yoshimura, K., Wang, L., Miralles, D.G., Jasechko, S., Lee, X., 2017. Revisiting the contribution of transpiration to global terrestrial evapotranspiration. Geophys. Res. Lett. https://doi.org/10.1002/2016GL072235

Willmott, C.J., 1981. On the validation of models. Phys. Geogr. https://doi.org/10.1080/02723646.1981.10642213

Wilson, K.B., Baldocchi, D.D., Hanson, P.J., 2001. Leaf age affects the seasonal pattern of photosynthetic capacity and net ecosystem exchange of carbon in a deciduous forest. Plant, Cell Environ. https://doi.org/10.1046/j.0016-8025.2001.00706.x

Wutzler, T., Lucas-Moffat, A., Migliavacca, M., Knauer, J., Sickel, K., Šigut, L., Menzer, O., Reichstein, M., 2018. Basic and extensible post-processing of eddy covariance flux data with REddyProc. Biogeosciences. https://doi.org/10.5194/bg-15-5015-2018

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Lettenmaier, D., Koren, V., Duan, Q., Mo, K., Fan, Y., Mocko, D., 2012. Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. J. Geophys. Res. Atmos. https://doi.org/10.1029/2011JD016048

Xu, T., Guo, Z., Liu, S., He, X., Meng, Y., Xu, Z., Xia, Y., Xiao, J., Zhang, Y., Ma, Y., Song, L., 2018. Evaluating Different Machine Learning Methods for Upscaling Evapotranspiration from Flux Towers to the Regional Scale. J. Geophys. Res. Amos. https://doi.org/10.1029/2018JD028447

Yang, F., White, M.A., Michaelis, A.R., Ichii, K., Hashimoto, H., Votava, P., Zhu, A.X., Nemani, R.R., 2006. Prediction of continental-scale evapotranspiration by combining MODIS and AmeriFlux data through support vector machine. IEEE Trans. Geosci. Remote Sens. https://doi.org/10.1109/TGRS.2006.876297

Yao, Y., Liang, S., Zhao, S., Zhang, Y., Qin, Q., Cheng, J., Jia, K., Xie, X., Zhang, N., Liu, M., 2013. Validation and application of the modified satellite-based Priestley-Taylor algorithm for mapping terrestrial evapotranspiration. Remote Sens. https://doi.org/10.3390/rs6010880

Yassin, M.A., Alazba, A.A., Mattar, M.A., 2016. Artificial neural networks versus gene expression programming for estimating reference evapotranspiration in arid climate. Agric. Water Manag. https://doi.org/10.1016/j.agwat.2015.09.009

Yin, J., Deng, Z., Ines, A.V.M., Wu, J., Rasu, E., 2020. Forecast of short-term daily reference evapotranspiration under limited meteorological variables using a hybrid bi-directional long short-term memory model (Bi-LSTM). Agric. Water Manag. https://doi.org/10.1016/j.agwat.2020.106386

Zaherpour, J., Mount, N., Gosling, S.N., Dankers, R., Eisner, S., Gerten, D., Liu, X., Masaki, Y., Müller Schmied, H., Tang, Q., Wada, Y., 2019. Exploring the value of machine learning for weighted multi-model combination of an ensemble of global hydrological models. Environ. Model. Softw. https://doi.org/10.1016/j.envsoft.2019.01.003

Zhao, P., Kang, S., Li, S., Ding, R., Tong, L., Du, T., 2018. Seasonal variations in vineyard ET partitioning and dual crop coefficients correlate with canopy development and surface soil moisture. Agric. Water Manag. https://doi.org/10.1016/j.agwat.2017.11.004

Zhao, T., Wang, Q.J., Schepen, A., Griffiths, M., 2019a. Ensemble forecasting of monthly and seasonal reference crop evapotranspiration based on global climate model outputs. Agric. For. Meteorol. https://doi.org/10.1016/j.agrformet.2018.10.001

Zhang, J., Zhu, Y., Zhang, X., Ye, M., Yang, J., 2018. Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2018.04.065

Zhao, W.L., Gentine, P., Reichstein, M., Zhang, Y., Zhou, S., Wen, Y., Lin, C., Li, X., Qiu, G.Y., 2019b. Physics-Constrained Machine Learning of Evapotranspiration. Geophys. Res. Lett. https://doi.org/10.1029/2019GL085291

Zou, L., Zhan, C., Xia, J., Wang, T., Gippel, C.J., 2017. Implementation of evapotranspiration data assimilation with catchment scale distributed hydrological model via an ensemble Kalman Filter. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2017.04.036

3.8 Supplemental Information

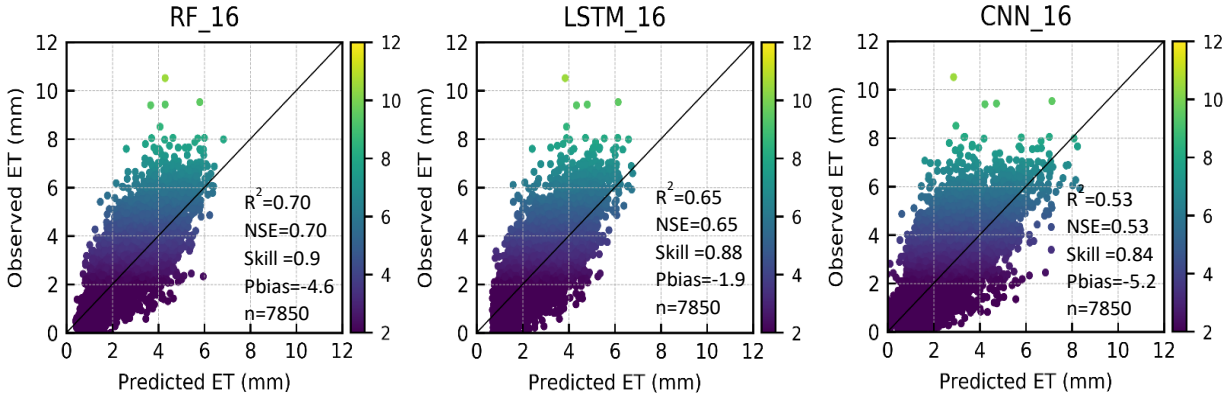

Figure S3.1: Comparison of RF_16, LSTM_16 and CNN_16 daily ET prediction models. Models are evaluated on 7850 daily ET observations (2003-2019) from different sites in Midwest. Model evaluation statistics are calculated on daily timestep. Here skill score is based on Wilmott's Index of Agreement.

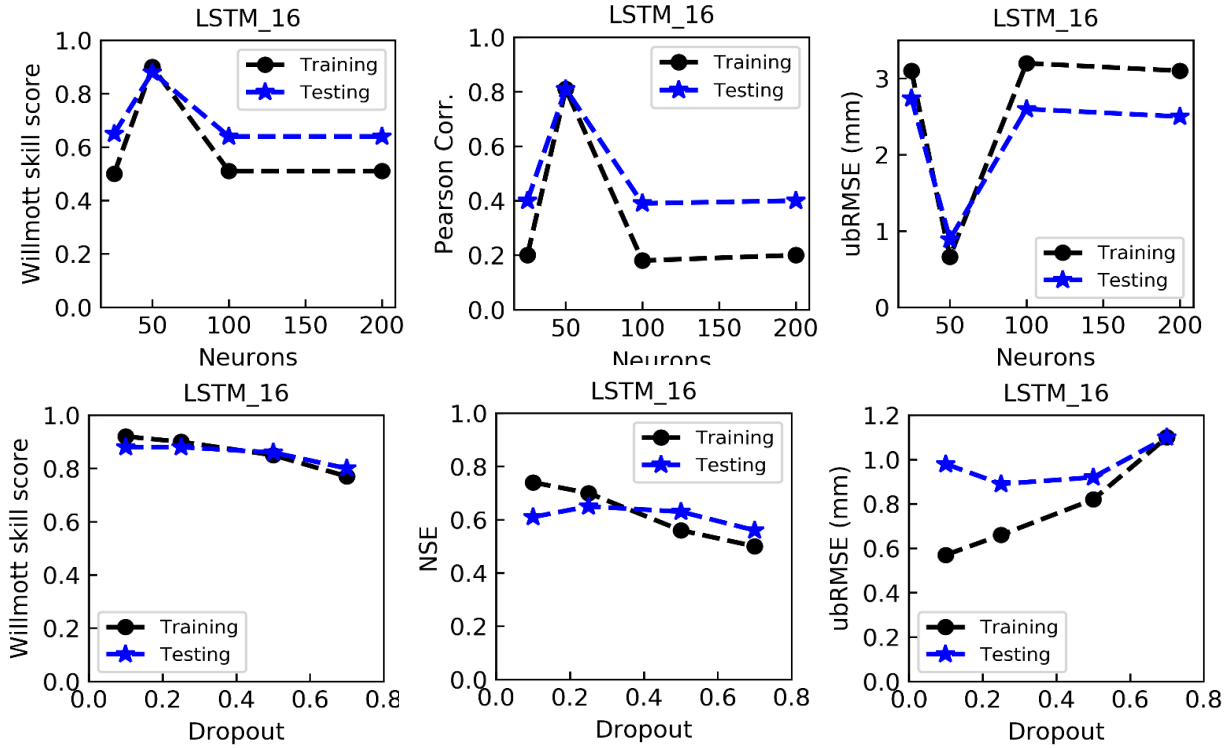

Figure S3.2: Selection of optimal number of neurons in layer one and dropout factor for LSTM_16. Evaluation of model performance on training and testing data set based on Willmott skill score, Pearson correlation and ubRMSE.

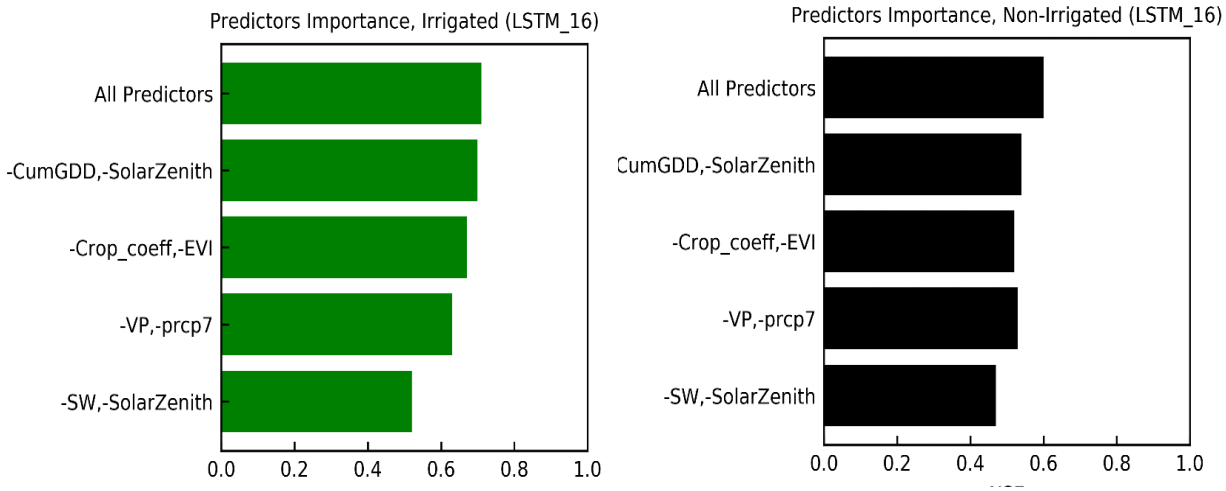

Figure S3.3: This figure shows predictors importance for LSTM_16 in irrigated and non-irrigated fields. Only most important predictors Here "–" sign means removal of a predictor. For example -SW,-SolarZenith '' means that these two predictors were removed from the model.

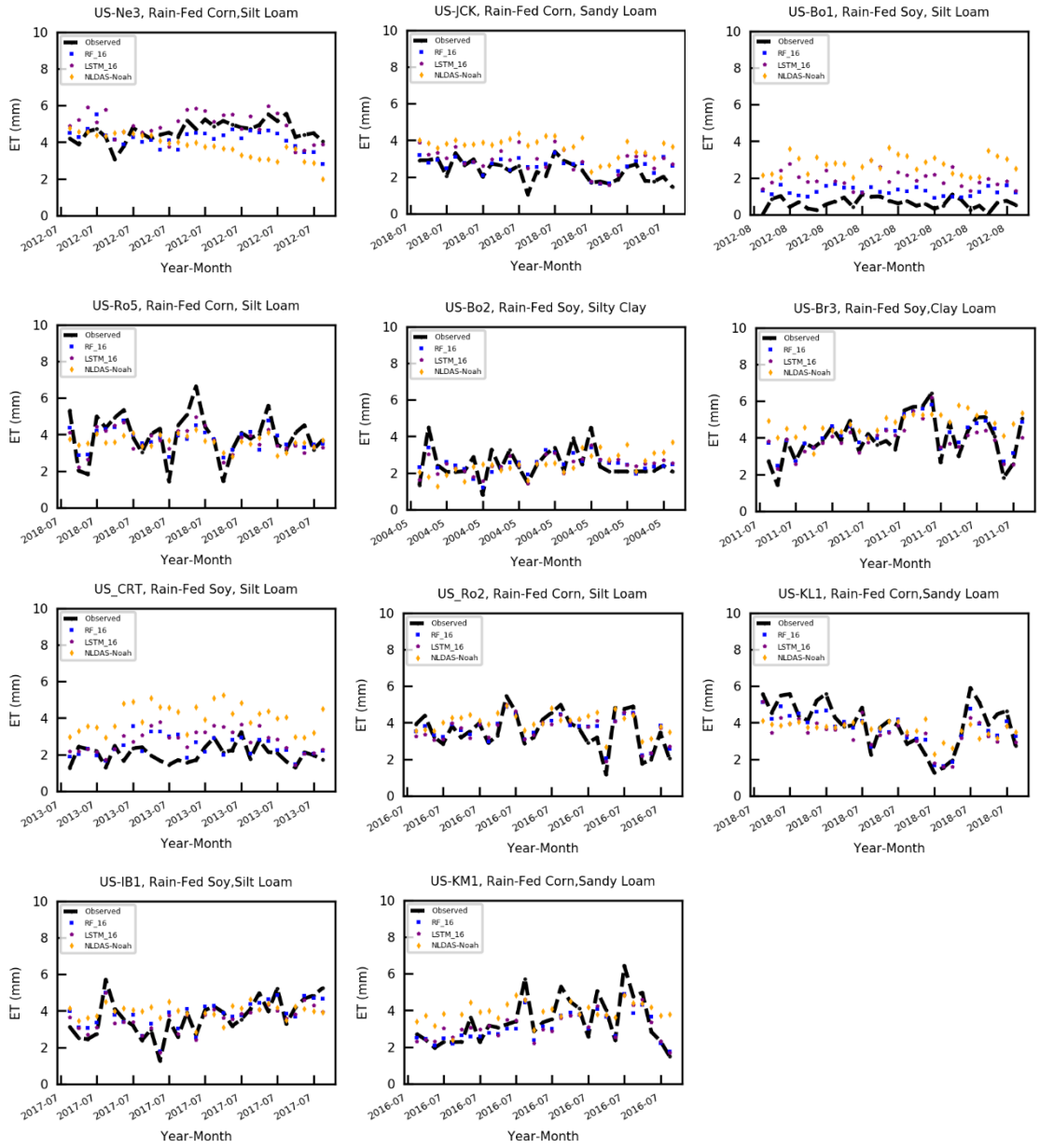

Figure S3.4: Rain-Fed sites used in training for RF_16, LSTM_16 prediction models. NLDAS_Noah is used as benchmark for comparison. Number at end of each model name shows the number of predictors used to build model. e.g RF_16 is RF model with 16 predictors and LSTM_16 is LTM model with 16 predictors.

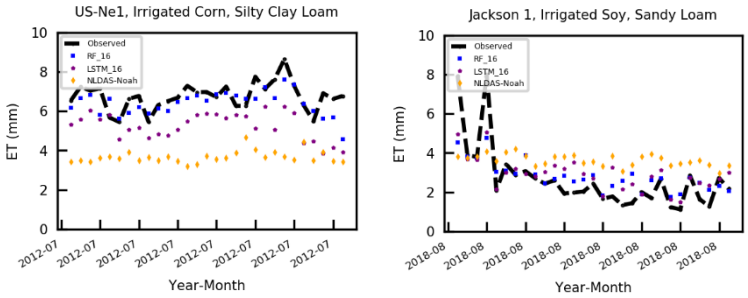

Figure S3.5: Irrigated sites used in training for RF_16,LSTM_16 prediction models. Model results are compared with NLDAS_Noah benchmark model comparison. Data is presented from sites located in MI (Jackson 1) and Ne (Us-Ne1).

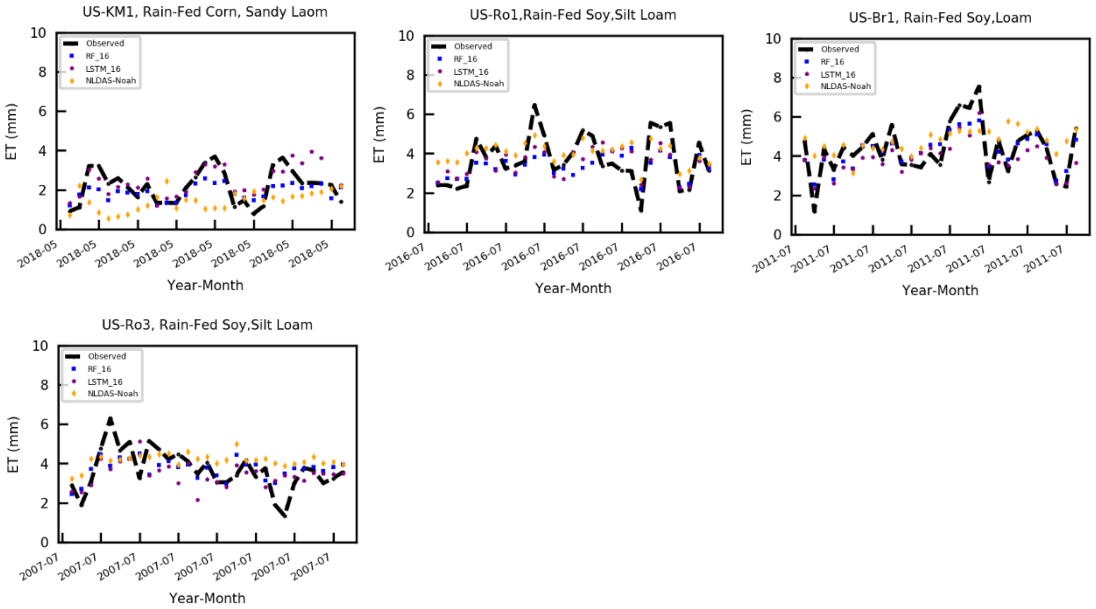

Figure S3.6:Rain-Fed Test sited used for evaluation of RF_16,LSTM_16 prediction models performance. NLDAS_Noah is used as benchmark for comparison. Data is presented from year 2007,2011, 2016 and 2018 across different sites.

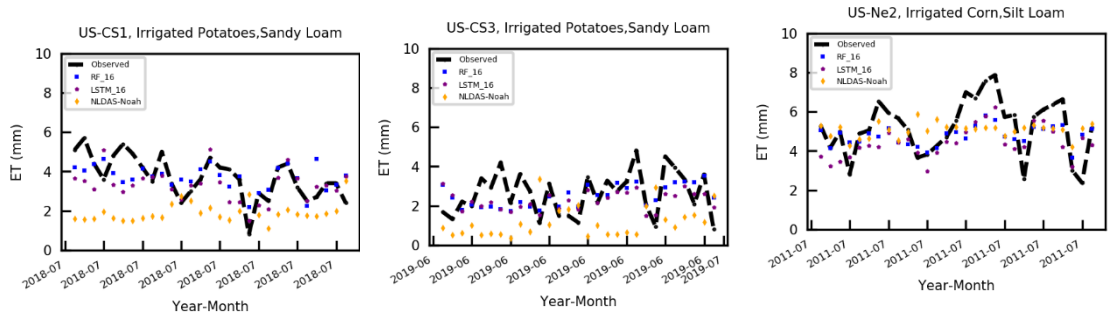

Figure S3.7: Irrigated test sited used for evaluation of RF_16,LSTM_16 prediction models performance in 2011, 2018 and 2019. NLDAS_Noah is used as benchmark for comparison.

Table S3.1: Comparison of RF_16, LSTM_16 and CNN_16 daily ET prediction models. Models are evaluated on 7850 daily ET observations (2003-2019) from seven test sites in Midwest. Model evaluation statistics are calculated on daily timestep.

Model Evaluation Statistics for prediction model	R^2	NSE	Willmott's skill score	Pearson Corr.	MAE (mm/day)	ubRMSE (mm/day)	RSR	pbias (%)	AIC
RF_16	0.70	0.70	0.90	0.84	0.64	0.75	0.55	-4.7	0.0
LSTM_16	0.65	0.65	0.88	0.81	0.72	0.89	0.59	-1.9	0.34
NLDAS_Noah	0.57	0.57	0.86	0.76	0.79	1.1	0.65	0.3	Benchmark
CNN_16	0.53	0.53	0.84	0.74	0.81	1.17	0.68	-5.2	0.93
RF_11	0.70	0.70	0.89	0.85	0.66	0.76	0.55	-5.3	0.04
LSTM_11	0.63	0.63	0.86	0.82	0.73	0.91	0.60	-6.0	0.42
CNN_11	0.63	0.63	0.86	0.82	0.73	0.92	0.61	-8.8	0.46
RF_5	0.63	0.63	0.85	0.81	0.73	0.94	0.61	-5.4	0.46
LSTM_5	0.53	0.53	0.80	0.75	0.82	1.20	0.69	-9.3	0.94
CNN_5	0.52	0.52	0.79	0.74	0.85	1.20	0.69	-5.5	1

Table S3.2: RF_16 prediction model evaluation by year (2003-2019) on seven test sites. Number of observations used in each year are shown with sample number (n). Evaluation metrices of R^2 NSE, Willmott's evaluation index or skill score, Pearson Correlation, ubRMSE, RSR and Pbias. are calculated on daily timestep.

Year	Site ID	Sample number (n)	R^2	NSE	Willmott's skill score	Pearson Corr.	MAE (mm/day)	ubRMSE (mm/day)	RSR	Pbias (%)
2003	US-Ne1	213	0.77	0.77	0.93	0.92	0.68	0.69	0.48	-15
2004	US-Ne2 US-Ro3 US-Ro1	523	0.57	0.57	0.86	0.76	0.66	0.80	0.66	2.9
2005	US-Ne2 US-Ro3 US-Br1 US-Ro1	832	0.70	0.7	0.90	0.87	0.69	0.93	0.54	-8.7
2006	US-Ne2 US-Ro3 US-Br1 US-Ro1	852	0.69	0.69	0.90	0.83	0.63	0.74	0.55	1.04
2007	US-Ne2 US-Ro3 US-Br1 US-Ro1	852	0.68	0.68	0.89	0.83	0.67	0.79	0.57	-2.2
2008	US-Ne2 US-Br1 US-Ro1	639	0.65	0.65	0.89	0.80	0.62	0.70	0.59	-1.63
2009	US-Ne2 US-Br1 US-Ro1	639	0.71	0.71	0.90	0.85	0.59	0.64	0.53	-6.0
2010	US-Ne2 US-Br1 US-Ro1	639	0.83	0.83	0.95	0.91	0.53	0.45	0.42	-3.0
2011	US-Ne2 US-Br1 US-Ro1	639	0.75	0.75	0.92	0.89	0.63	0.74	0.5	-9.7
2012	US-Ne2 US-Ro1	426	0.77	0.77	0.93	0.89	0.66	0.79	0.48	-6.2
2013	US-Ne2 US-Ro1	273	0.50	0.50	0.80	0.82	0.77	1.1	0.70	-21.0
2014	US-Ro1	213	0.55	0.55	0.84	0.75	0.61	0.64	0.67	-4.3
2015	US-Ro1	213	0.56	0.56	0.82	0.87	0.76	1.02	0.66	-19.4
2016	US-Ro1	213	0.64	0.64	0.86	0.8	0.65	0.67	0.60	2.6
2017	MI-KM1	213	0.63	0.63	0.87	0.8	0.54	0.48	0.6	-3.4
2018	US-KL1 US-CS1	337	0.66	0.66	0.89	0.82	0.54	0.54	0.58	2.8
2019	US-CS3	99	0.34	0.34	0.76	0.70	0.81	0.77	0.81	26.7

Chapter 4

4. Spatial and Temporal Forecasting of Groundwater anomalies

Abstract

Monitoring anomalies in groundwater (GW) over long periods of time is essential to understanding groundwater resource variability. However, it is challenging to predict GW anomalies over the long term in agricultural areas due to complicated boundary conditions, heterogeneous hydrogeological characteristics, and groundwater extraction, as well as nonlinear interactions among these factors. To overcome this challenge, here, we developed an advanced modeling framework based on a recurrent neural network of long short-term memory (LSTM) as an alternative to complex and computationally expensive physical models. The objective of this study was to forecast GW anomalies two months in advance and to evaluate the drivers that influence GW dynamics in densely irrigated agricultural regions. An application and evaluation of this new approach was conducted in the Wisconsin Central Sands (WCS) region in the U.S., one of the most productive agricultural regions. The modeling framework was developed for the time period of 1958-2020 by utilizing easily accessible dynamic and static variables to represent hydrometeorological and geological characteristics. Additionally, GW fluctuation data was utilized from 26 piezometers (wells) implanted in the sandy aquifer in WCS over a period of 10-60 years. The subset of GW observations from ~10-60 years, not used in model training, can forecast GW anomalies two months out with a coefficient of determination R^2 ~ 0.8. Additionally, MAE was

less than 0.35 m month^{-1} across the study region for both temporal and spatial modeling [3]frameworks. According to this study, groundwater anomalies showed high spatiotemporal variability, and their responses are influenced by boundary condition, catchment geology, climate, and topography differently across locations. Additionally, land use change and irrigation pumping have interactive effects on GW anomalies forecasting. The results of our modeling framework suggest that our framework can be used as an alternative method of simulating water availability and groundwater level changes in areas where subsurface properties are unknown or difficult to determine.

Keywords

Machine learning, Groundwater anomalies, drivers, irrigation pumping

Key points

- With machine learning methods, anomalies in depth to groundwater can be forecast two months in advance.
- Due to their high spatiotemporal variability, catchment geology, climate, topography, and boundary condition have a strong but variable and site-specific influence on long-term groundwater changes
- Irrigation pumping has a stronger influence on groundwater forecasting than land use change.

[3] Chapter 4 has been submitted in the Journal of Hydrology for review (HYDROL53214)

4.1 Introduction

Globally, groundwater (GW) is responsible for about 30% of freshwater reserves (Schneider et al., 2011), provides about half of the drinking water (Majumdar et al., 2020), and is a critical component of the water-food-energy nexus (Luyssaert et al., 2014; Al-Yaari et al., 2022). However, even though groundwater is a valuable resource, societies are increasingly interfering with it by expanding irrigated crop fields (Siebert et al., 2015), which can exacerbate hydrologic extremes like floods and droughts (Milly et al., 2002). Water use in irrigation, for instance, accounts for 70% of the world's freshwater withdrawals, and irrigated areas have grown proportionately faster than the world's population. Aside from its effects on agricultural productivity (Troy et al., 2015; Smidt et al., 2016; Deines et al., 2017; Nie, 2022), irrigation also modulates local to regional atmospheric processes and land–atmosphere interactions (Pokhrel et al., 2012), alters precipitation patterns by altering partitioning of the surface energy balance, and through biogeochemical and biogeophysical effects on climate (DeAngelis et al., 2010; Lu et al., 2017; Hurtt et al., 2020).

To monitor the changes in GW dynamics, many modeling studies have been undertaken to represent the human water footprint so as to describe hydrological and biogeochemical processes better (Wada et al., 2017). However, several hydrological variables affect groundwater dynamics, such as streamflow, precipitation, snowmelt, and system boundary conditions, climate variability and GW pumping (Hintze et al., 2020; Cai et al., 2021). It is, therefore, challenging to

create groundwater-level simulations. Generally, traditional 3-D process-based models of groundwater levels based on physical processes of water movement require computationally-intensive mathematical solutions and extensive basin attribute knowledge (Cuthbert et al., 2019). Groundwater systems have nonlinear interactions, spatial heterogeneity, and time lags that are difficult to describe, and as a result, conventional process-based groundwater level models have been criticized as overly complicated and difficult to apply (Clark et al., 2015).

A large region can be monitored using remote sensing techniques owing to a contiguous coverage of satellite data (Leidner & Buchanan 2019). For assessing different quantities associated with the global water cycle, there is a multitude of space-borne missions that provide satellite products. Among these products, terrestrial ET has been quantified by MODIS (Moderate Resolution Imaging Spectroradiometer), precipitation has been quantified by GPM (Global Precipitation Measurement), soil moisture has been quantified by SMAP (Soil Moisture Active Passive), and land use has been quantified by USDA-NASS (United States Department of Agriculture-National Agricultural Statistics Service) (MardanDoost et al., 2019). Increasingly, remote sensing estimates of land use are used to estimate irrigated agricultural areas based on land use data sets derived from remote sensing. However, the resolution of these products is too coarse for local estimations of GW flux, and they have not been directly linked to forecasts of spatial region GW anomalies and drivers.

Satellite data and process-based models have limitations, so water resources models increasingly combine observed data with computational experiments to provide a foundation for

developing theories related to system processes (Zahura et al., 2012, Lary et al., 2016, Chen et al., 2020; Hu et al., 2021), particularly those pertaining to agricultural water withdrawal, whose influence is not sufficiently explained by existing theory (Clark et al., 2015, Karpatne et al., 2017; Ekblad & Herman, 2021). Furthermore, data that describes these processes in multiple sectors makes it possible to infer plausible models from them as a complement to the approach. Therefore, an empirical solution for nonlinear and nonparametric relationships can be found through machine learning (ML), a form of artificial intelligence. Models based on machine learning, such as artificial neural networks (ANNs), have been successfully applied to predict nonlinear hydrological processes such as precipitation, runoff, stream flow, and water quality (CITATION?). Although the physical meanings of the simulation results are not taken into consideration during the machine learning process, their performance can often be explained by multiple physical relationships explained through the learning process (Brodrick et al., 2019). In hydrology, variations in variables are often closely related to the physical characteristics of catchments. The main recharge methods for groundwater level in the continental United States (CONUS) have been studied in previous studies based on the different regional characteristics of the recharge methods (Gleeson et al., 2012), indicating that machine learning can be used to examine how drainage basin characteristics and groundwater fluctuations affect results.

Groundwater levels can also be simulated using machine learning-based hydrological models (Sudheer et al., 2002; Krishna et al. 2008; Yoon et al., 2011; Liu and Shi, 2019; Malekzadeh et al., 2019). Convolutional neural networks, long short-term memory (LSTM), and genetic programming are some examples of advanced, recurrent neural networks (RNNs) (Kratzert et al.,

2019). However, despite the fact that ML has been used by hydrologists for decades to address different hydrologic problems, it is still in its infancy as many long-standing and unconventional sources of data remain primarily unexplored and fragmented, and incomplete information remains substantial and diverse (Bergen et al., 2019). In addition, despite research on many of these water cycle components, remote sensing data sets are rarely combined in densely irrigated areas to forecast spatial GW anomalies and evaluate drivers, which limits the ability to estimate these fluxes to specific circumstances.

Most methods for estimating GW fluctuations rely on estimates of agricultural water demand, such as evapotranspiration (ET) and soil models, together with estimates of surface water availability. Additionally, a variety of factors, including dynamic climate variability and precipitation, as well as crop rotations, contribute to significant local interannual variability in irrigated areas (Wisser et al., 2008; Döll et al., 2012; Brown & Pervez, 2014). An example of the approach considering these factors, is the MODFLOW-based Little Plover model (Bradbury, et al. 2017). Based on land use data and an ET model that considers temperature, precipitation, crop type, and root depth, the Little Plover River model has been used to simulate GW dynamics in Wisconsin Central Sands (WCS). Additionally, the model allocates the remaining water demand to GW pumping using known surface water availability. These models are beneficial and have successfully been used (Kraft et al., 2016) but calibrating them over large geographic areas is time-consuming and expensive due to the number of parameters and ground-truth data needed. Moreover**,** relatively few studies have focused on characterizing and forecasting the interannual variations of GW water anomalies across large spatial regions. At this scale, research is limited

on how meteorological and human factors affect long-term ground water anomalies since projected changes in meteorological variables vary across regions, with different hydrological systems responding differently to those changes. It may therefore be important to explore useful information about regional drivers of GW fluxes by combining local GW monitoring data with multispectral satellite estimates and measured pumping data.

To build the groundwork for that challenge, here, in this study, we rely on an RNN, a state-of-the-art machine learning algorithm, to combine remote sensing datasets from the subsurface, ground, and space to better understand the interconnection between the climate, biophysical drivers, and human activities in the form of land use change and pumping in WCS to forecast GW anomalies two months in advance over WCS region. We especially ask 1) How can easy-to-obtain remote sensing derived environmental datasets be combined to accurately forecast GW anomalies across a densely irrigated regions, using computationally efficient ML models, 2) How do GW anomalies forecast vary in space and across the WCS, and 3) How can irrigation pumping, and crop type change influence this anomaly forecasting. In this paper, we seek a data-driven approach that utilizes satellite data to forecast GW anomalies at the local scale and accomplish this by integrating diverse satellite data sets related to different water balance components. As a result of this investigation, we will have a better understanding of aquifers' physical mechanisms, natural variability, and response. An outcome of this work is to support water managers and food security efforts to have spatial monthly GW anomalies forecast results that accurately delineate agricultural areas under stress, give insight into factors influencing GW

fluctuation behavior, and clarify how climate change and biophysical parameters affect GW dynamics.

4.2 Methods

4.2.1 Description of study area

Wisconsin Central Sands (WCS), our study area (Fig. 4.1), comprises 630,000 ha of land bounded by the Fox and Wolf River basins on the east and the Wisconsin River on the west. A coarse-grained, glacially drifted sand aquifer typically >30 m thick and a water table 4.3 to 20 m below the surface characterize the WCS region hydrogeologically (Nocco et al., 2018). The WCS region utilizes over 2,100 high-capacity wells (Fig. 4.1c) to water over 86,500 ha of potato, corn (field and sweet), peas, and snap beans (Keene & Mitchell, 2010), supporting a market worth US$550 million in the northern Great Lakes region (Smail et al., 2016).

Growers in the WCS rely on high-capacity wells to pump groundwater from the unconfined aquifer, which nourishes 1,000 km of trout streams and 80 lakes (Kraft et al., 2012). In the late 1960s, when over 700 high-capacity wells were in the WCS, pumping already affected water levels and streamflows (Weeks & Stangland, 1971). Despite this, irrigated land use conversion continued through the 2000s, and by the mid-2000s, more than 3,000 high-capacity wells had been constructed in the WCS. During a period of modestly dry weather, severe pumping impacts were observed to streams, lakes, and wetlands in the region. As a result, stretches of the Little Plover River, a Class I trout stream with recreation significance to aquatic stakeholders in the WCS, dried out over a period of five years from 2005 to 2009 (Kraft & Mechenich, 2010). The

USGS conducted studies of groundwater-surface water interactions in the early 1960s at this site and found that due to groundwater declines, surface water stress, and irrigation expansion, the region's water resources are under pressure.

Average annual precipitation in the WCS ranges from 790 to 810 mm, divided between 30 to 40% recharge and 60 to 70% ET (Kucharik et al., 2010, Nocco et al., 2019). Regional climate studies report that there has been an increase in annual precipitation of 100 to 200 millimeters between 1948 and 2007, with 50 to 100 millimeters between June and August, with 70% of the precipitation partitioned to groundwater recharge during the past 60 years under native potential vegetation (Motew & Kucharik, 2013). Because groundwater recharge should have increased without conversion to irrigated land cover, climate change cannot explain the observed declines in groundwater and surface water stress in the WCS. Instead, the observed hydrological stresses and groundwater level declines across the WCS region in the 2000s are attributed to a 45 mm reduction in annual recharge (equivalent to net consumed irrigation) on irrigated lands (Kraft et al., 2012). Due to the irrigated systems, there is a corresponding increase in ET that causes a reduction in recharge. Therefore, since 2000, particularly from 2005 to 2009, there has been a renewed interest in irrigation diversions and the consequent lowering of water levels in the central sands.

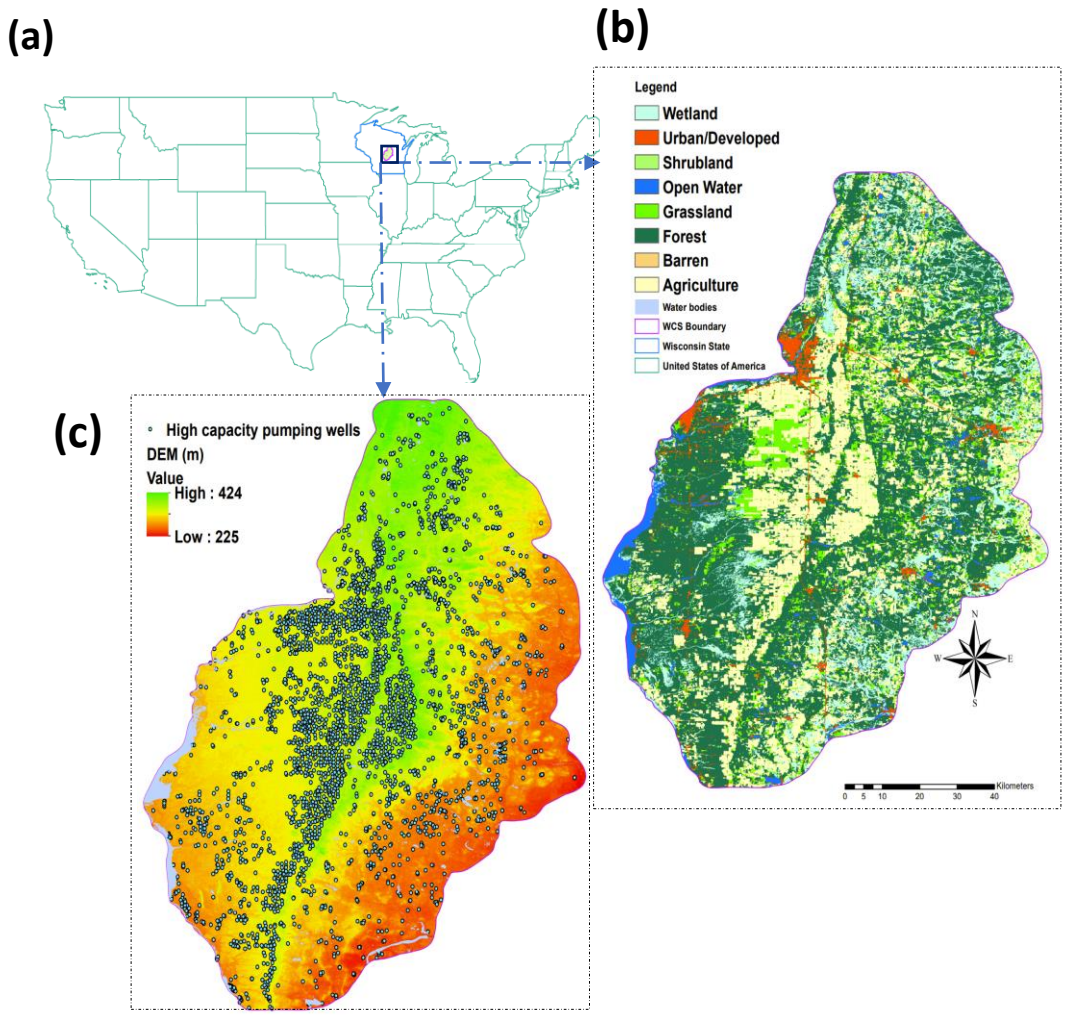

Figure 4.1: WCS region is presented. (a) Location of Wisconsin central sands (WCS) in the USA, (b) Land use is shown along with water bodies, (c) High-capacity pumping wells and DEM of the region.

4.2.2 Input data:

For GW anomalies forecasting, static and dynamic data have been used.

4.2.2.1 Dynamic, temporal data

To represent the characteristics of the study area in our proposed model, fourteen dynamic climatic water balance variables for the time period 1958-2020 were obtained from

TerraClimate (https://developers.google.com/earth-engine/datasets/catalog/IDAHO_EPSCOR_TERRACLIMATE Abatzoglou et al., 2018). TerraClimate is a monthly climatic water balance dataset for global terrestrial surfaces from 1958-present. These data provide important inputs for ecological and hydrological studies at global scales that require high spatial resolution and time-varying data. TerraClimate variables have monthly temporal resolution and a ~4-km (1/24th degree) spatial resolution. Out of fourteen climatic water balance dynamic variables (Table 1), PCA was done on eleven meteorological-related variables to reduce dimensionality, while the remaining three water balance-related variables, such as runoff, snow water equivalent (SWE), and soil moisture, were used without PCA. The eleven meteorological variables for PCA include actual evapotranspiration, climate water deficit, Palmer drought severity index (PDSI), reference evapotranspiration, precipitation, minimum temperature, maximum temperature, downward surface shortwave radiation, vapor pressure, vapor pressure deficit, and wind speed. After performing PCA on ten meteorological variables, seven PCAs with combined eigenvalues greater than 95% were selected for model inputs.

USGS GW monitoring well data from 1958 to 2020 (https://water.usgs.gov/watuse/data/), were aggregated on a monthly basis. Twenty-six wells were selected in WCS because well measurements in other locations have insufficient points (< 120 months). Water withdrawal from 2010 to 2020 (dnr.wisconsin.gov/topic/WaterUse/WithdrawalSummary.html) was used (Table 1). USGS modeled Land Cover Land Use (LCLU) (sciencebase.gov/catalog/item/59d3c73de4b05fe04cc3d1d1) was used for the time period of 2010-2020. That time period was selected to match pumping data, which was only available from

2010-2020.

4.2.2.2 Spatial, static data

An Automated Geoscientific Analyses (SAGA) Geographic Information System (GIS) software (https://saga-gis.sourceforge.io/saga_tool_doc/2.2.3/a2z.html) was provided with digital elevation map (DEM) data from United States Geological Survey (USGS) (10 m) to calculate twenty primary and secondary topographical terrain and morphometry variables such as elevation, slope, aspect, mean curvature (meanc), plan curvature (planc), profile curvature (profc), tangential curvature (tangc), and surface roughness factor (roughness) from analytical hillshading, and distance to channel and water bodies (Olaya, 2009; Csillik & Drăguţ, 2018) to represent the spatial characteristics of our study area (Table 4.1). PCA was performed on these twenty topographic terrain and morphometry variables to extract the first fourteen PCAs with eigenvalues that explained more than 95% of the variance. The physical and chemical properties of soil affect water retention and redistribution in space and time (Mohanty and Skaggs, 2001). This study used 250-m resolution maps of soil properties to extract three variables, including clay, sand content, and bulk density (BD), from the SoilGrids (Hengl et al., 2017). Two geology-related variables were extracted from maps of geological units (mrdata.usgs.gov/geology/state/state.php?state=WI). These datasets have been widely used by others to map GW potential across the world (e.g., Naghibi et al., 2016; Kalantar et al., 2019; Martínez-Santos & Renard, 2020) and were utilized as input data for the ML models in the WCS. Soils and geology variables were included as inputs without PCA.

Table 4.1: Descriptions and sources of static and dynamic data are presented. Details about the different variables used in each category and their sources are also described.

Variable Type	Variable name	Resolution	Description	Source
Climate or meteorology (Dynamic, temporal data)	actual evapotranspiration (AET), climate water deficit, Palmer Drought Severity Index (PDSI), reference evapotranspiration (PET), precipitation, minimum temperature, maximum temperature, downward surface shortwave radiation, vapor pressure, vapor pressure deficit, wind speed	4.6 (km), monthly	AET calculated using a Thornthwaite-Mather climatic water-balance model (Willmott et al., 1985), Climate water deficit is defined as the difference between PET and AET, PSDI is a standardized index based on a simplified soil water balance and estimates relative soil moisture conditions, PET is aalculated using ASCE Penman-Montieth	TerraClimate (Abatzoglou, 2018)
Water balance (Dynamic, temporal data)	Runoff, soil moisture, snow water equivalent (SWE)		Runoff, soil moisture and SWE derived using a one-dimensional soil water balance model	TerraClimate (Abatzoglou, Williams, et al., 2018)
GW monitoring well data	Depth to GW (m)	Daily (26 wells)	Monitoring well data from 1958 to 2020, aggregated to a monthly basis	https://water.usgs.gov/watuse/data/
	Water withdrawal data	monthly	pumping data for irrigation from 2011 to 2020 was obtained from DNR WI.	dnr.wisconsin.gov/topic/WaterUse/WithdrawalSummary.html
Land Cover Land Us	USGS modeled Land Cover Land Use (LCLU)	yearly	USGS modeled LCLU was used for time period 2010-2020	sciencebase.gov/catalog/item/59d3c73de4b05fe04cc3d1d1
Topography & morphometry (Spatial, static data)	elevation, slope, aspect, mean curvature (meanc), plan curvature (planc), profile curvature (profc), tangential curvature (tangc), surface roughness factor (roughness), analytical hillshading, , convergence index (CI), closed depressions (CD), drainage basin, topographical wetness index (TWI), slope length (LS) factor, channel network, channel network base level (CNBL), distance to Channel Network (VDCN), valley depth (VD), relative slope position (RSP), distance to water bodies	10 (m)	Digital elevation map was used to calculate topographic and morphometry variables using GIS SAGA.	Olaya, 2009; Csillik & Drăguț, 2018
Soil (Spatial, static data)	Sand, clay, bulk density	250 (m)	Soil properties such as clay, sand and bulk density	Ramcharan et al., (2018)
Geology (Spatial, static data)	Rock type		maps of geological units and faults	

4.2.2.3 Timeseries Analysis

The total observation for the long-term GW depth measurement from twenty-six wells for time period 1958-2020 are n= 8126. The long-term GW depth measurements in the WCS were used to calculate depth to GW anomalies by subtracting the depth to GW from the long term mean depth to GW for each well. Additionally, autocorrelation of the previous two-month GW anomalies was computed (Table 4.2). Mean depth the GW was measured to assess whether a well is shallow or deep well. If the depth to GW was greater than five meters, then it was called deep GW well. Based on that threshold we had ten deep wells and sixteen shallow wells. Maximum and minimum and mean GW anomalies were also calculated. This analysis was done to compare the patterns of GW dynamics across different well locations and evaluate the location-specific and time-dependent effects of topography, geological units, and soil properties on GW dynamics.

Table 4.2: Characteristics of long-term GW monitoring wells are described. Well labels are generated based on the last 4-5 unique numbers of the USGS well ID. To be concise, well labels were used whenever wells were described.

Well Label	USGS Well ID	Longitude	Latitude	Time Period	Sample Size (n)	Autocorrelation (AC)	Mean depth to GW (m)	Minimum anomalies	Maximum anomalies
df_0001	ID_435758089490001	-89.817	43.966	1958-1970	149	0.90	4.6	-0.77	0.87
df_0801	ID_442822089290801	-89.486	44.473	1959-1968	104	0.45	1.9	-0.37	0.31
df_1101	ID_441414089091101	-89.153	44.237	1958-1994	435	0.92	11.1	-1.21	1.24
df_1201	ID_434014089471201	-89.787	43.671	1970-1996	322	0.93	12.3	-1.38	1.01
df_1601	ID_434924089191601	-89.321	43.823	1959-1963	55	0.78	10.0	-0.51	0.58
df_1701	ID_441531089291701	-89.488	44.259	1964-1974	116	0.86	3.6	-1.56	0.96
df_2101	ID_441250089312101	-89.523	44.214	1958-1969	142	0.76	4.4	-0.90	0.88
df_2201	ID_441236089272201	-89.456	44.210	1968-1979	184	0.84	4.4	-1.11	0.89
df_2601	ID_442619088592601	-88.991	44.439	1958-1967	112	0.85	6.2	-0.65	0.90

df_2701	ID_442623089302701	-89.508	44.440	1959-2006	555	0.82	3.9	-2.40	1.12
df_2801	ID_441454089432801	-89.725	44.248	1964-1993	348	0.59	2.8	-2.07	0.88
df_2901	ID_442622089302901	-89.508	44.440	2006-2020	165	0.85	3.7	-1.28	1.04
df_3001	ID_441452089433001	-89.725	44.248	1995-2020	310	0.64	3.0	-1.66	0.92
df_3401	ID_435244089293401	-89.493	43.879	1958-2020	754	0.87	4.7	-1.03	1.27
df_04101	ID_441806089104101	-89.178	44.302	1958-1979	250	0.81	6.3	-0.94	0.45
df_4101	ID_443127089174101	-89.295	44.524	1958-1978	252	0.98	10.3	-1.35	1.80
df_4201	ID_443126089174201	-89.295	44.524	2010-2020	120	0.92	9.1	-1.42	0.91
df_4301	ID_442313089474301	-89.799	44.388	1958-1981	274	0.52	1.9	-0.47	0.41
df_4501	ID_442810089194501	-89.333	44.471	1958-2020	748	0.92	1.5	-2.35	1.91
df_4701	ID_441448089384701	-89.647	44.247	1958-1982	292	0.47	1.2	-1.05	0.96
df_4901	ID_442100089384901	-89.647	44.350	1958-1982	295	0.42	1.3	-0.89	0.79
df_5101	ID_442610089385101	-89.648	44.436	1958-1968	130	0.78	5.1	-0.69	0.66
df_5301	ID_444709089265301	-89.448	44.786	1958-2003	552	0.96	5.8	-2.03	2.19
df_5601	ID_441833089315601	-89.529	44.310	1958-2020	752	0.48	1.6	-1.31	0.68
df_5701	ID_440721089315701	-89.533	44.122	1958-1966	101	0.95	6.0	-0.91	0.87
df_490001	ID_435759089490001	-89.817	43.967	1969-2020	609	0.67	4.4	-1.40	1.09

4.4.3 Model Development

In this section, first an overview of LSTM was presented and then the development of models specific to this study was explained.

4.4.3.1 Overview of LSTM model

Known for their predictive ability and capability to analyze sequential patterns, recurrent neural networks are a specific type of neural network (Zewdie et al., 2021; Rybalkin et al., 2020). An experimental study conducted by Tan et al. (2022) has demonstrated that recurrent neural networks work on the principle of cyclical connectivity between neurons and information flow between neurons. Since they have loops across hidden layers, previous information can be linked to the current task, allowing information to persist. Recurrent neural networks are more powerful

than ordinary feed-forward networks because of these cyclic connections. Therefore, a recurrent neural network's outcome depends on the results of the previous computations since a recurrent neural network has "memory" that encodes information from the previous computations. (Graves, 2012).

In recent years, the Long-Short Term Memory (LSTM) network has gained popularity and is applied to solving scientific problems (Asanjan et al, 2018). The LSTM improves recurrent neural networks by eliminating the vanishing gradient problem that inhibits recurrent neural networks from learning (Graves, 2012, Houdt et al., 2020). In a long sequence recurrent neural network, the gradient of the loss function is highly reduced during backpropagation time during the vanishing gradient phase (Bengio et al., 1994). An LSTM avoids vanishing gradients by storing information for an extended period through the interaction between its four repeating units. The four structures are called the cell state or the memory part of the LSTM unit, and the three other structures are known as the input, output, and forget gates. Memory gates comprise the input gate, the tanh neural network layer, and a bitwise multiplication operation. The forget gate is based on a sigmoid neural network and a bitwise multiplication. Using the tanh function and bitwise multiplication, the output gate transmits the state and signal from the cell to the output (Blender & Fraedrich et al., 2006).

4.4.3.2 Two month forecast GW anomaly models

The temporal model for GW anomalies forecast two months in advance was built using static and dynamic variables. GW anomalies can inform us about the direct change in GW

compared to GW elevation. A temporal model can also be used to assess the transferability of models across time periods. Although the input environmental data have different spatial and temporal resolutions, previous studies have shown that it is feasible to integrate these heterogeneous datasets (e.g., coarse-scale climate data with fine-scale land surface parameters) for modeling soil moisture and water fluxes (e.g., soil moisture, runoff, evapotranspiration) at fine spatial resolutions (Jiang et al., 2020; Vergopolan et al., 2020). The temporal model was developed by using the input of seven meteorological-related PCAs, three water balance-related variables (SWE, soil moisture, and runoff), fourteen PCAs from topographic & morphometry-related variables, three variables for soil properties and two geology-related variables, and boundary condition of previous two-month GW anomalies. The model output for temporal variables was two months out GW anomalies.

The model fitting was done with Python version 3.9 software. The temporal model was developed for the time period 1958-2020 and was calibrated using the first part of the time series (~70%) from twenty-six wells (Fig. 4.2a). The time period for GW monitoring data was different for different wells. The temporal model was calibrated or trained on 70% of data from the time period of 1958-2017 using n=5684 observations from twenty-six wells. Calibrated model parameters are presented in Table 4.3. After calibration, the model was tested on the same twenty-six wells but with different observations (n=2442) from the time period of 1963-2020. Evaluation of model performance was based on the coefficient of determination (R^2), Nash–Sutcliffe model efficiency (NSE), root mean square error (RMSE), and mean absolute error (MAE).

In addition to the temporal model, a spatial model was also developed to study the

transferability of the model across wells at different locations within the WCS. The spatial model was trained on 70% of data using 6,036 points from the time period of 1958-2020 taken from eighteen wells (Fig. 4.2b). Not all wells had data for the same time period. The spatial model was then tested on a sample of 2,090 observations taken from eight different wells not used in the training but in the WCS.

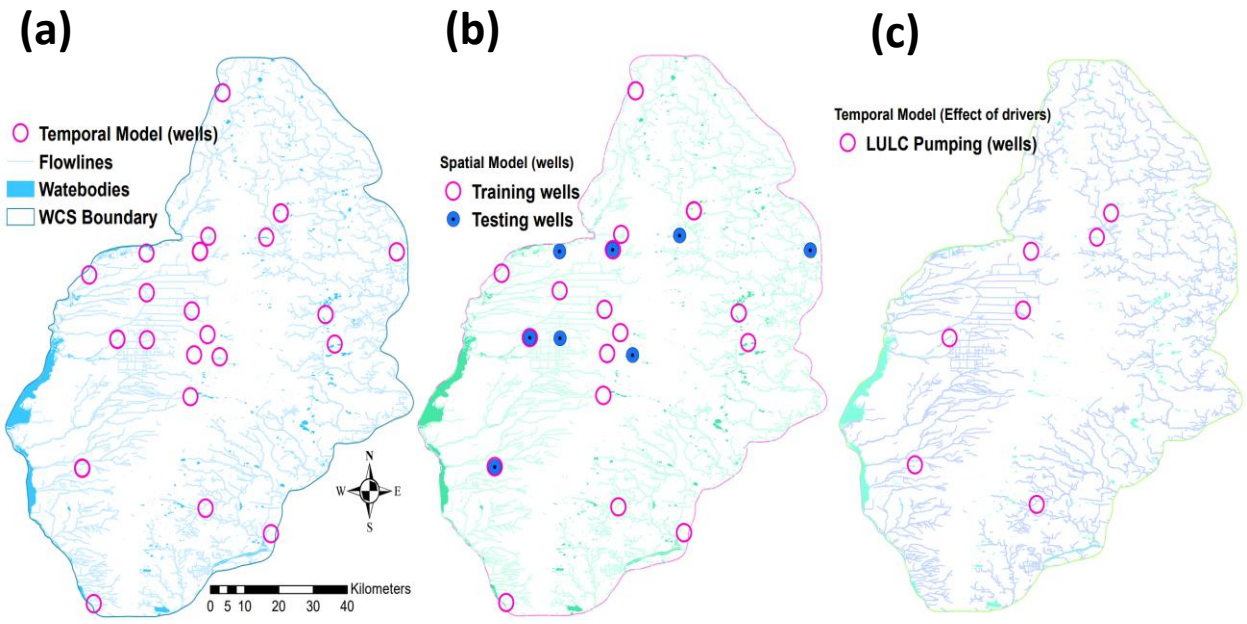

Figure 4.2: Location of training and testing wells for GW anomalies forecast models.
(a) Temporal model, the same 26 wells were used for calibration and testing but using different parts of time series, calibration or training time period: 1958-2017 (n=5684), testing time period: 1963-2020 (n=2442)
(b) Spatial model. Eighteen wells were used for calibration, time period 1958-2020 (n=6036), and eight different wells were used for testing, time period 1958-2020 (n=2090).
(c) A temporal model was developed to study the influence of land use and pumping. The same seven wells were used for calibration and testing but using different parts of time series, calibration or training time period: 2010-2017 (n=574), testing time period: 2016-2020 (n=242)

4.4.4 Models to study the influence of drivers on GW anomalies forecast.

The influence of different drivers was tested by excluding and including drivers from the temporal model. Since the goal was to assess the drivers for all wells, the results from the temporal model were used. The effect of drivers was studied by excluding driver categories (e.g., meteorology-related variables, soil-related variables, etc.), and then calculating the percentage change in MAE was computed to assess the influence of those drivers. Hence, the model without SWE variables had 29 input variables, and the model without topography & morphometry-related variables had 16 input variables. The influence of meteorology and soils was also examined. Twenty-three input variables were included in the model to study the effect of climatic and soil-related variables. Another model was built using 29 input variables after removing the boundary condition of the previous two-month anomalies. For the temporal model built to study the influence of land use and pumping, only five topography & morphometry-related PCAs were included. Hence the total input variables used in the land use and pumping model were based on 21 input variables after considering climatic, soil, land use, and pumping-related variables.

A temporal model was also developed to study the influence of land use change and irrigation pumping. Using the initial part of the time series (2010-2017, n=574), the temporal model was trained on seven wells, then tested on the same seven wells (Fig. 4.2c) but using data from a different period (2016-2020, n=242). At different wells, GW monitoring data was collected over different periods of time. Because irrigation pumping data was unavailable in WCS until 2010, a shorter time period was selected. By systematically removing land use and pumping variables, the effect of land use and pumping was assessed.

Table 4.3: The structure of different models developed in this study is shown. Information about variables, sample size, and calibrated model parameters is also presented.

	Number of input variables	Number of wells (training)	Testing sample (n)	Number of wells (testing)	LSTM layer 1 neurons	LSTM layer 1 drop out	LSTM layer 1 neurons	LSTM layer 1 drop out
Temporal model	30	26	2442	18	300	0.5	200	80
Spatial Model	30	18	2090	8	300	0.5	200	80
Temporal model (land use, irrigation pumping influence)	21	7	242	7	200	0.5	100	80

4.3 Results

4.3.1 Temporal model performance for two-month GW anomalies forecast

The overall performance of the two-month temporal forecast model is shown in Fig. 4.3. We obtained an overall training score of R^2= 0.91. Moreover, error metrics suggest we achieved good prediction results for the testing data set, R^2 = 0.84, RMSE = 0.22 m, and MAE = 0.16 m. For the training data set, no significant difference was found between model performance and depth to GW anomalies forecast for shallow and deep GW anomalies. However, the difference between R^2 (p = 0.0035) and MAE (p = 0.0049) of shallow and deep GW anomalies was significant. For testing data, MAE and RMSE of forecasted anomalies for shallow GW were higher than deep GW.

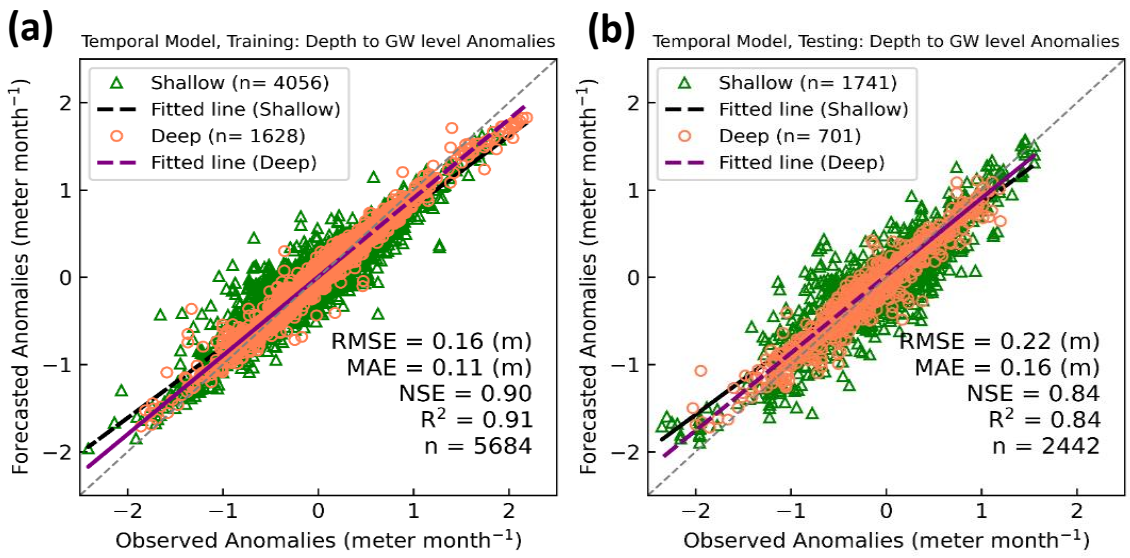

Figure 4.3: Observed and two months out forecasted depth to GW level anomalies for temporal model (a) model performance for training data (period 1958-2017), (b) testing data (period 1963-2020). Data were marked to distinguish between shallow (green) versus deep (orange) GW wells. While evaluation statistics were calculated for shallow and deep GW anomalies together, fitted lines are shown separately for shallow vs. deep GW anomalies. Total sample size is also presented. Additionally sample size is also shown separately for the shallow and deep GW wells.

The MAE for testing data ranged from 0.07-0.24 m (Fig. 4a). MAE performance of the model was not always correlated with larger sample sizes (Table 4.4). Wells with a larger sample size (n) did not consistently yield better results. For example, well_4501 MAE was 0.21 despite having a higher autocorrelation (AC=0.92) and being trained on a larger sample size (n=748), compared to well_2601 (n=112, A =0.85). Well_2601 had a smaller MAE of 0.079 m and a lower RMSE of 0.1 m in the testing data. The RMSE ranged from 0.1-0.32 m. The wells with the highest MAE and RMSE, while not clustered at one location, were located along the midline of the WCS regions (Fig. 4.4).

The R^2 across various locations in WCS ranged from 0.33-0.96 for testing data (Fig.3c). Wells with R^2 less than 0.75 had an average depth to GW of 3.9 m (15 wells), while wells with R^2 greater than 0.75 had an average depth to GW of 6.6 m (11 wells). Generally, wells with high AC

for anomalies have higher R^2 and more accurate forecasts. But there were some exceptions; for example, well_3401 has AC of 0.87 with previous two-month anomalies, which is higher than autocorrelation for well_04101 (Table 4.4). However, well_3401 R^2 of 0.66 is lower than R^2 of 0.81 for well_04101. Moreover, well_4501 has a greater AC of 0.92 and a larger MAE of 0.2 m compared to well_0801, which had a lower AC of 0.45 and a lower MAE of 0.082 m. The R^2 for the wells was not inversely correlated with their MAE and RMSE. For example, well_4501 has a higher R^2 of 0.92 (AC=0.92) and a higher MAE =0.21 m and RMSE= 0.29 m.

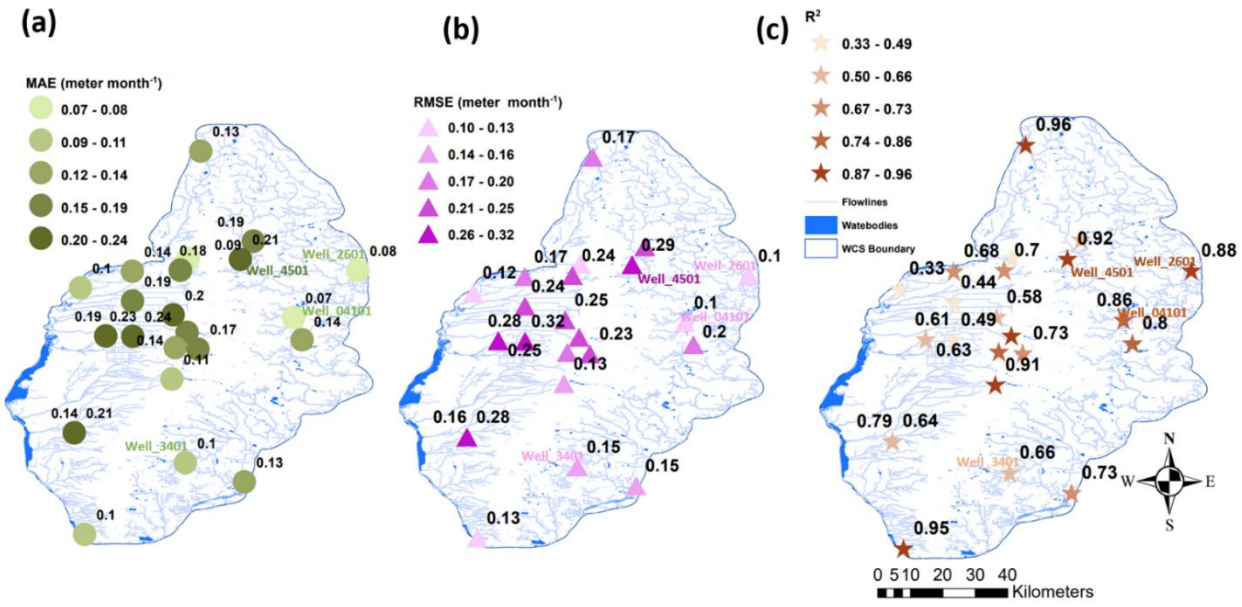

Figure 4.4: Performance of a two-month out temporal forecast model for 26 wells on testing data. Evaluation metrics are shown (a) MAE, (b) RMSE, and (c) R^2 as well. The map also indicates flowlines, waterbodies, and the boundary of the WCS. To be concise, the last four-five unique digits of USGS codes show the labels of some of the wells. Labels of the wells belong to the well marker they are touching.

The time series plot for best simulation performance (Fig. 4.5a), worst simulation performance (Fig. 4.5b), and moderate simulation performance (Fig. 4.5c) based on R^2 and NSE

are shown. Generally, a correlation was observed between model R^2 and the variance of the wells (σ^2) (Table 4.4). For example, among these three wells, the best-performing well_4501 has the highest R^2=0.92 and σ^2 =0.78 for observed GW anomalies (Fig. 4.5a) compared to well_4301 with the lowest R^2=0.33 and lowest σ^2 of 0.027 for observed GW anomalies. However, R^2 and MAE were not correlated. For example, because of smaller σ^2, the MAE of worst R^2 well_3401 was smaller than well_4501 (MAE: .095 vs. 0.207). Overall, the model was able to capture the sub-seasonal variations.

Table 4.4. Evaluation metrics for the two-month forecast temporal model are shown for the testing data. Time period and sample size are shown for both calibration and testing data. Autocorrelation and depth to GW anomalies variance is calculated based on both training and testing periods.

Well Label	R^2	NSE	MAE (m month^{-1})	RMSE (m month^{-1})	Calibration Period	Calibration sample size (n)	Testing Period	Autocorrelation (AC)	Anomalies variance (σ^2)
df_0001	0.79	0.42	0.14	0.16	1958-1966	104	1966-1970	0.90	0.16
df_0801	0.48	0.03	0.08	0.11	1959-1965	73	1966-1968	0.45	0.02
df_1101	0.80	0.78	0.15	0.20	1958-1983	304	1983-1994	0.92	0.25
df_1201	0.95	0.94	0.10	0.13	1970-1988	225	1988-1996	0.93	0.32
df_1601	0.73	0.56	0.13	0.15	1959-1962	38	1962-1963	0.78	0.09
df_1701	0.88	0.75	0.19	0.25	1964-1971	81	1971-1974	0.86	0.30
df_2101	0.79	0.73	0.14	0.19	1958-1966	99	1966-1969	0.76	0.21
df_2201	0.73	0.55	0.17	0.23	1964-1975	129	1975-1979	0.84	0.26
df_2601	0.88	0.85	0.08	0.10	1958-1964	78	1965-1967	0.85	0.15
df_2701	0.70	0.65	0.17	0.22	1959-1992	388	1992-2006	0.82	0.24
df_2801	0.63	0.59	0.19	0.25	1964-1984	244	1984-1993	0.59	0.19
df_2901	0.72	0.70	0.18	0.24	2006-2016	115	2016-2020	0.85	0.35
df_3001	0.61	0.61	0.22	0.28	1995-2013	217	2013-2020	0.64	0.21
df_3401	0.66	0.62	0.10	0.15	1958-2001	528	2002-2020	0.87	0.12
df_04101	0.86	0.85	0.07	0.10	1958-1972	175	1972-1979	0.81	0.05
df_4101	0.96	0.95	0.09	0.11	1958-1972	176	1972-1978	0.98	0.45
df_4201	0.66	0.64	0.19	0.22	2010-2017	84	2017-2020	0.92	0.37
df_4301	0.33	0.05	0.09	0.12	1958-1974	192	1974-1981	0.52	0.03
df_4501	0.92	0.92	0.21	0.29	1958-2002	524	2002-2020	0.92	0.78
df_4701	0.49	0.47	0.24	0.32	1958-1974	204	1975-1982	0.47	0.19
df_4901	0.44	-0.01	0.19	0.25	1958-1975	206	1975-1982	0.42	0.06

df_5101	0.68	0.65	0.14	0.17	1958-1965	91	1965-1968	0.78	0.11
df_5301	0.96	0.95	0.12	0.17	1958-1990	386	1990-2003	0.96	0.69
df_5601	0.58	0.47	0.20	0.25	1958-2001	526	2001-2020	0.48	0.10
df_5701	0.91	0.98	0.11	0.13	1958-1963	71	1963-1966	0.95	0.27
df_490001	0.64	0.63	0.21	0.28	1969-2005	426	2005-2020	0.67	0.18

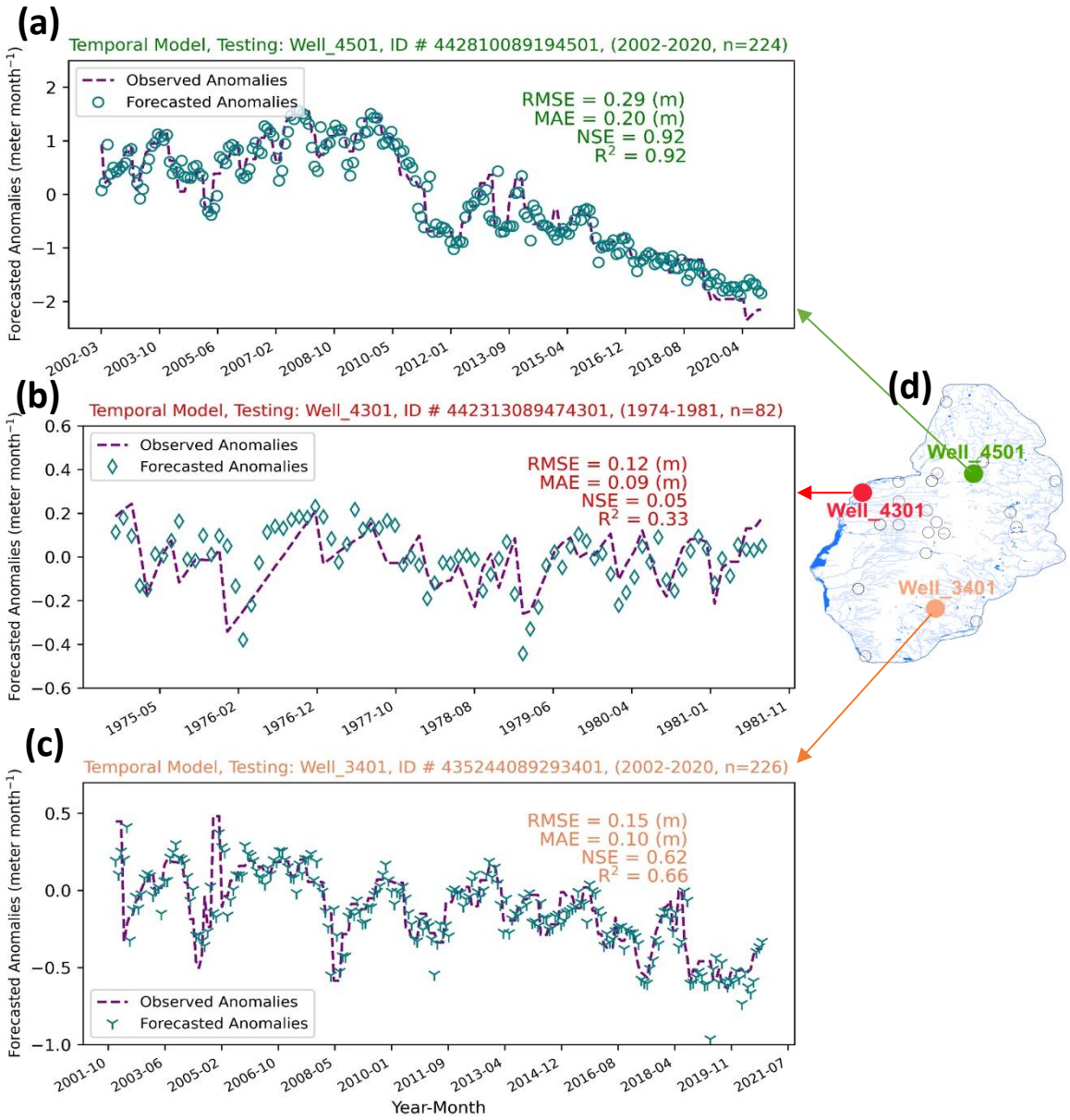

Figure 4.5: Temporal model performance two months in advance in terms of R^2 and NSE for testing data, (a) best performance well (calibration period: 1958-2002, n=524), (b) worst performance well (calibration period: 1958-1974, n=192) and (c) moderate performance well (calibration period: 1958-2001, n=528). (d) The location of these wells is also shown on the map. Evaluation metrics of RMSE and MAE are also shown.

4.3.2 Spatial model performance for two-month GW anomalies forecast

Fig. 4.6 shows the performance of the spatial model for training (a) and testing data (b). Overall, the coefficient of determination R^2 is 0.89, and the RMSE is 0.16 m for the training data. The R^2 is 0.79, and the RMSE is 0.33 m for the testing data. The model performance was not significantly different for deep GW anomalies for shallow vs. deep water.

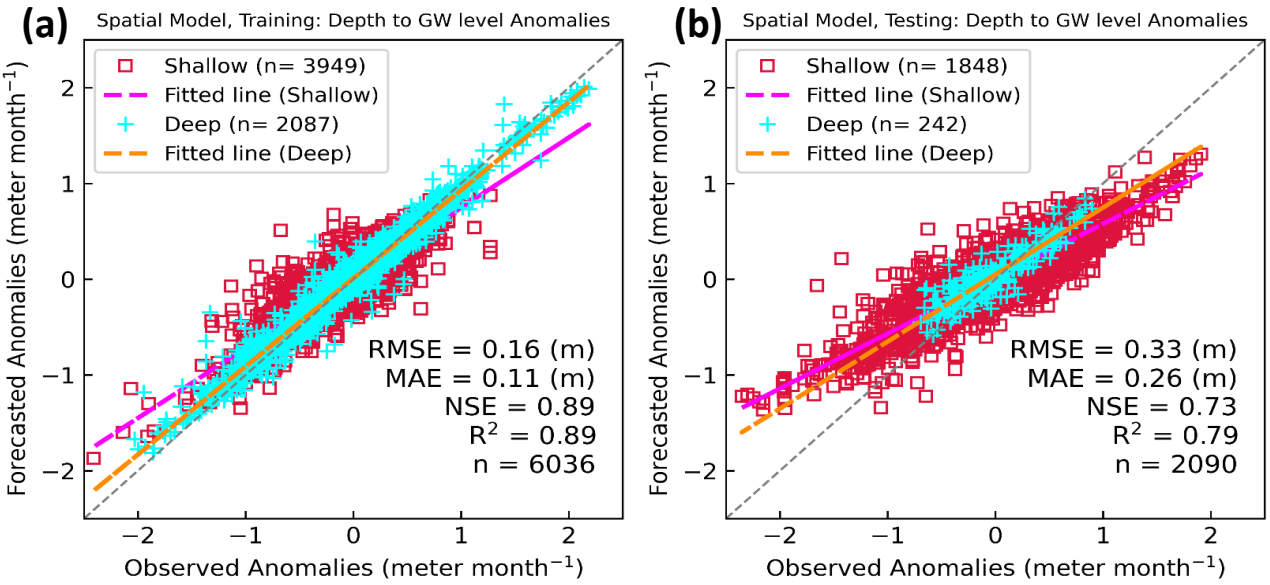

Figure 4.6: Performance of the spatial model (a) training data (time period: 1958-2020, number of wells=18), (b) testing data (time period: 1958-2020, number of wells=8). Model evaluation statistics are also presented. Each well has a different time period. Sample size is shown separately for the shallow and deep GW wells. Additionally total sample size is also presented. Fitted lines are shown separately for shallow vs. deep GW anomalies. Evaluation statistics were calculated for shallow and deep GW anomalies together.

The MAE range for testing locations was 0.16-0.41 m (Fig. 4.7a). In the testing data, there were two wells with deep GW and six wells with shallow GW. While overall there was no significant difference between model performance for shallow and deep wells, the lowest MAE

of 0.12 m and 0.14 m belonged to two wells with deeper GW, i.e., well_2601 and well_5101, respectively (Table 4.5). Generally, wells with higher AC had higher R^2. However, MAE and RMSE were not related to AC. For example, well_4701 had a lower AC=0.47 compared to well_4501 (AC=0.92) but had a lower MAE of 0.25 compared to 0.34 for well_4501. Additionally, MAE has a correlation of 0.82 with observed GW anomalies σ2 (Table 4.5). That is why with the increase in GW anomalies σ2, MAE also increased. For example, the highest variance of 0.78 and the highest MAE=0.34 m $month^{-1}$ and RMSE=0.41 m $month^{-1}$ were found for well_4501. RMSE was also strongly correlated (corr=0.88) with minimum anomalies. For example, well_4501, with the largest minimum anomalies of -2.3 m, also had the largest RMSE of 0.41 m $month^{-1}$ but had the highest R^2 of 0.90. The R^2 range for testing wells was 0.50-0.90 (Fig. 4.7c). Similarly, the RMSE was correlated (corr=0.77) with maximum anomalies and tended to get larger with larger anomalies.

Fig. 4.8 shows time series plots of testing wells. Well_4501 performed best on R^2 and NSE. However, the same well also had the highest MAE and RMSE. The spatial model captured the aspects of the signal in GW anomalies, and the time series plots helped identify misfits. For example, the model underestimated the negative anomalies for spring 2002 and 2004 for well_3001. In addition, some extreme positive anomalies in March 1965 and negative anomalies in early spring 1974 were missed for well_2201 (Fig. 4.8c). Overall, the model accurately characterized the sub-seasonal variations of signals and forecasted trends with plausible shapes.

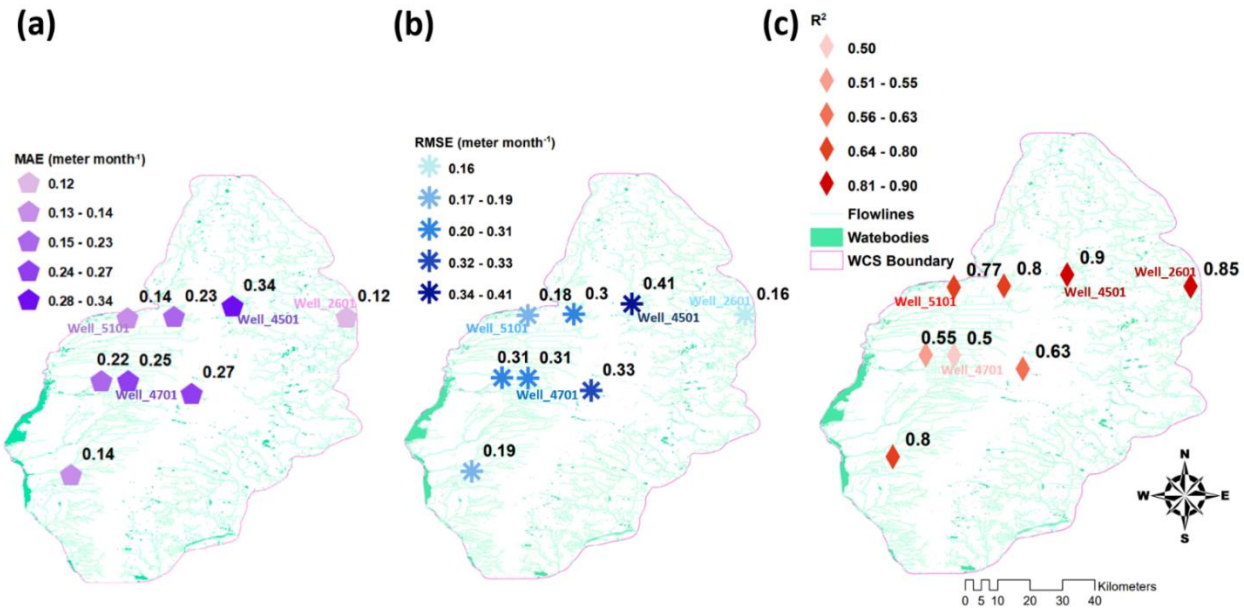

Fig. 4.7: Spatial two-month forecast model for depth to GW anomalies. Evaluation metrics for testing wells (a) MAE, (b) RMSE, and (c) R^2 are shown. Spatial model was calibrated on eighteen wells (period: 1958-2020, n=3949) and tested on eight wells (period: 1958-2020, n=1848). Only a few wells are labeled to be concise. Well labels belong to the well marker, they are touching.

Table 4.5. Performance of spatial two month out forecast model for testing data. Evaluation statistics of R^2, NSE, MAE and RMSE are included along with other characteristics of well

Well Label	R^2	NSE	MAE (m month^{-1})	RMSE (m month^{-1})	Testing Period	Testing sample size	AC	Mean depth to GW (m)	Anomalies variance σ^2
df_0001	0.80	0.78	0.14	0.19	1958-1970	149	0.90	4.60	0.16
df_2201	0.63	0.58	0.27	0.33	1968-1979	184	0.84	4.42	0.26
df_2601	0.85	0.83	0.12	0.16	1958-1967	112	0.85	6.23	0.15
df_2901	0.80	0.74	0.23	0.30	2006-2020	165	0.85	3.72	0.35
df_3001	0.55	0.54	0.22	0.31	1995-2020	310	0.64	2.97	0.21
df_4501	0.90	0.79	0.34	0.41	1958-2020	748	0.92	1.48	0.78
df_4701	0.50	0.50	0.25	0.31	1958-1982	292	0.47	1.24	0.19
df_5101	0.77	0.72	0.14	0.18	1958-1968	130	0.78	5.07	0.11

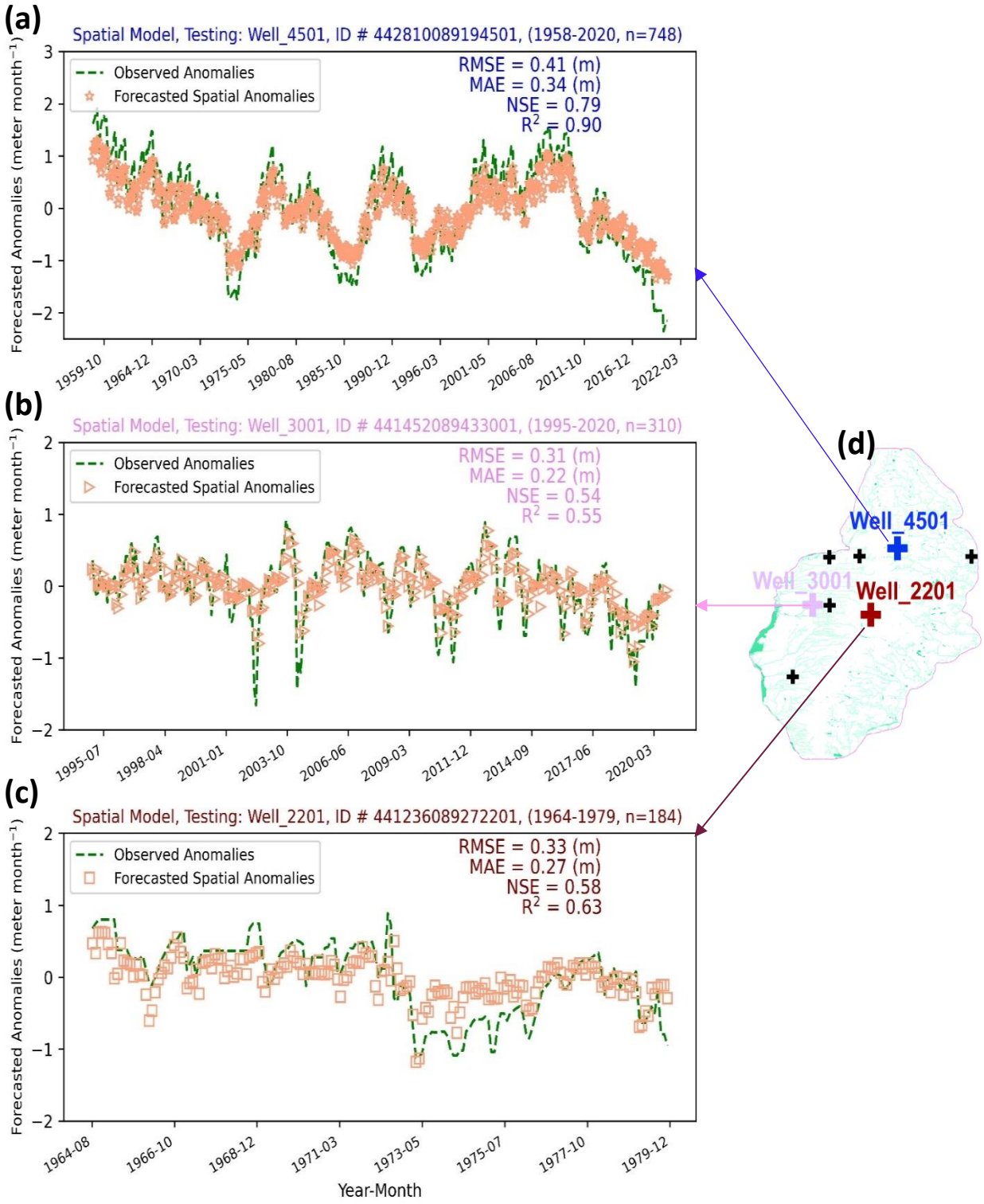

Figure 4.8: Times series plot for testing data (a) best performance well_4501, (b) poor performance well_3001, (c) moderate performance well_2201 based on R^2 and NSE, (d) location of wells is shown on the map.

4.3.3 Comparison of temporal and spatial forecast model

The performance of two-month GW anomalies was compared between temporal and spatial models (Fig. 4.9). The R^2 range was 0.33-0.96 for the temporal model and 0.50-0.90 for the spatial model. Overall, the R^2 of the spatial model was higher than the temporal model, implying strong extrapolation performance. Similarly, the NSE for the spatial model was higher than the temporal model. But the temporal model NSE has a larger range. Contrary to the pattern observed for R^2 and NSE, the spatial model has a larger range for MAE and RMSE. For example, MAE was 0.12-0.34 for the spatial model and 0.074-0.24 m month^{-1} for the temporal model. Similarly, RMSE ranged from 0.16-0.41 m for the spatial model and 0.098-0.319 m for the temporal model. Hence overall temporal model has a lower mean R^2, and NSE with a wider range. A spatial model, on the other hand, had a greater MAE and RMSE with a wide range.

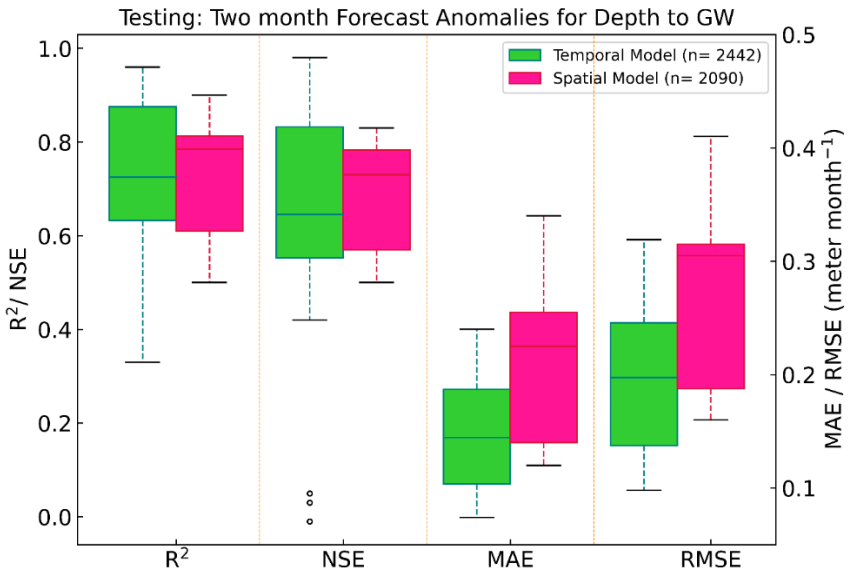

Figure 4.9: As a measure of R^2, NSE, MAE, and RMSE, a box plot is shown for comparing spatial and temporal two-month forecast models. The temporal model was tested on 26 wells but with fewer data points because the first part of the time series was used for calibration. The spatial model was tested on eight wells, but the whole time series from those wells were included in the testing because the spatial model was calibrated on the remaining eighteen wells.

4.3.4 Effect of drivers on GW anomalies forecast

The effect of drivers on depth to GW anomalies forecast is shown in Fig. 4.10 in terms of percentage (%) change in MAE. By excluding SWE, the percentage change in MAE was from -32 to 46% (range=78, Fig. 4.10b), for ten wells. For example, the two wells where MAE decreased more than 10% after removing SWE include well_0001 (% change in MAE = -32%) and well_4301(% change in MAE = -11%). On the other hand, there were four wells where the percent change in MAE increased by more than 10% (Table 4.6). For example, two wells with the highest MAE increase include well_1701, with a 20% MAE increase, and well_4101, with a 46% MAE increase. The effect of topography on model performance was also evaluated (Fig. 4.10c). The % change in MAE after excluding topography-related drivers was from -54 to 55% (range=110). Hence, surprisingly, the effect of topography-related drivers is more conspicuous than SWE drivers. For example, for four wells, MAE was reduced by greater than 10%, but for eight wells, MAE increased by greater than 10%. Among those eight wells, MAE increased to 43% for well_4101, 44% for well_1101, and 55% for well_04101 after removing topography-associated drivers. Additionally, these three wells with highest % change in MAE were located in northeast side of WCS.

When meteorology-related drivers were removed from the model (Fig. 4.10d), the percent change in MAE was from -41 to 72 % (range 114). Hence the effect of excluding meteorological drivers was more significant than excluding SWE or topography-associated drivers. Following excluding meteorology-related drivers, MAE was reduced by more than 10% for four wells and increased by more than 10% for eight wells. Moreover, for well_1701 (deep

GW), the MAE increase was 72% (Table 4.6). When drivers related to soil characteristics were removed from the model (Fig. 4.10e), the % change in MAE was from -39-79% (range 118). For example, MAE was reduced by more than 10% for two wells, and for eight wells, MAE was increased by more than 10%. For well_4101, the MAE increase was 79%. Excluding boundary conditions from the model had the most conspicuous impact on MAE (Fig. 4.10f). For example, MAE changed from 4 to 644% (range=640). Hence for all wells except well_5101, the MAE error was greater than 10%. For 16 wells, the MAE increased by more than 100%. All drivers, except meteorology-associated drivers, increased MAE for well_4101 when excluded. Well_1701 MAE was also increased when other drivers were excluded except for topography-related drivers.

4.3.5: Land use change and GW pumping effect

Fig. 4.11 shows the effect of land use and pumping drivers on the two-month forecast GW anomalies temporal model. R^2 was increased from 0.41 to 0.45 (9.2% improvement), and NSE was improved from 0.20 to 0.23 (14.1% improvement) by incorporating information about land use changes into the model (Fig. 4.11a). By including pumping information, NSE was improved 30% from the baseline model. Hence pumping seems to influence GW anomalies forecasts more than land use. The R^2 was further augmented by including information about land use change and GW pumping in the same model (Fig. 4.11d). Compared to the baseline model, the R^2 was enhanced from 0.41 to 0.53, and MAE was reduced from 0.29 to 0.26 m. Additionally, NSE improved by 77.8% from the baseline model change, and RMSE was reduced to 10.3% (Fig. 4.12b).

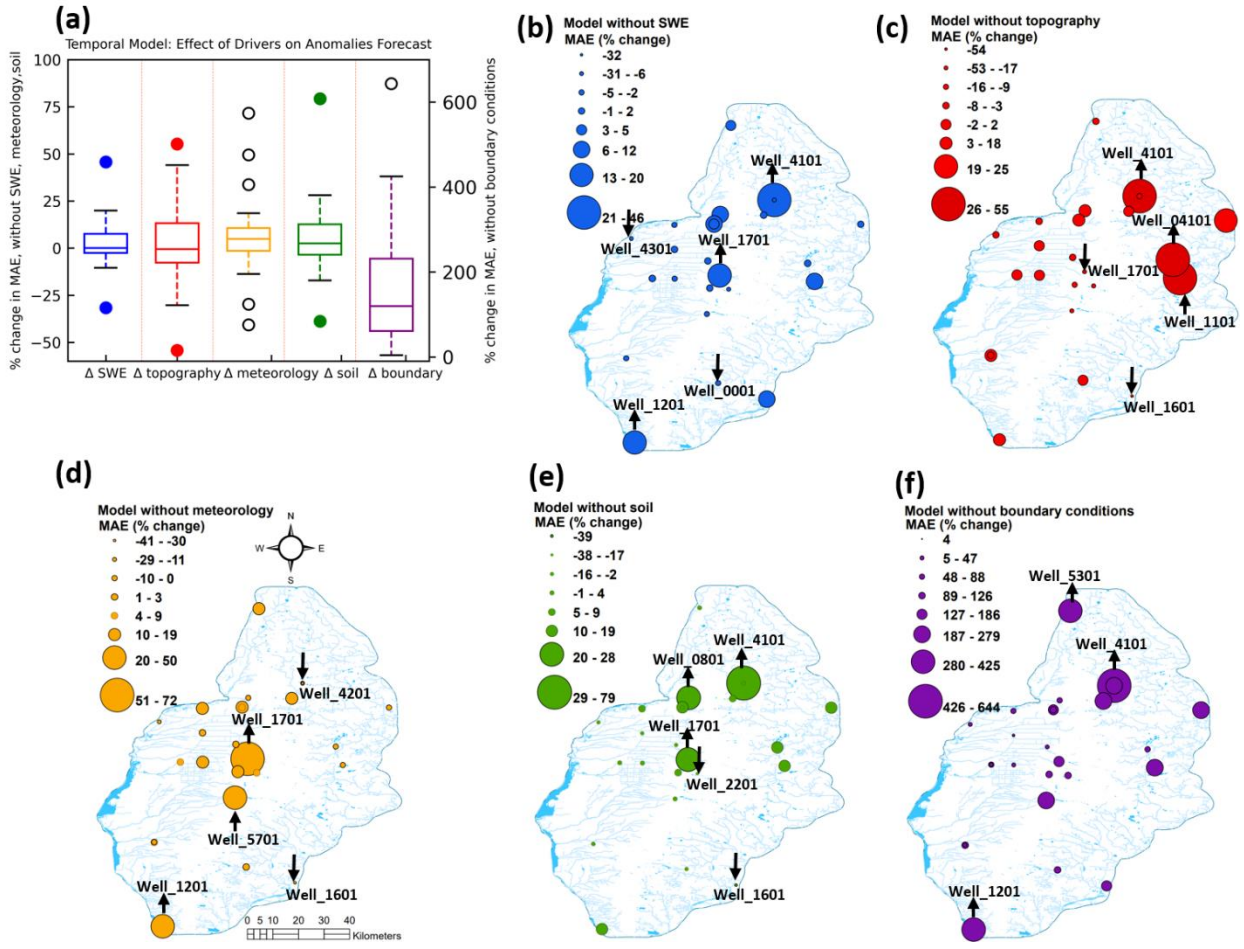

Figure 4.10: The effect of drivers on the performance of the temporal model for testing data, (a) Percentage (%) change in MAE after excluding drivers from the model, (b) % change in MAE after excluding SWE, (c) % change in MAE after excluding drivers related to soil, (d) % change in MAE after excluding meteorology related drivers, (e) % change in MAE after excluding topography related drivers, and (f) % change in MAE after excluding boundary conditions. A positive % change value means an increase in MAE, and a negative % change value means a decrease in % change. The upward arrow is used to show an increase in % MAE, and the downward arrow is used to show a decrease in % MAE. Only wells with remarkable change are labeled and marked by arrows.

Table 4.6. Percentage change in MAE due to exclusion (Δ) of drivers from temporal model is presented for the testing data. Positive value mean increase in MAE and negative value mean decrease in MAE. Characteristics of the wells such as mean depth to GW, autocorrelation (AC)

Well Label	Δ SWE MAE (%)	Δ Topography MAE (%)	Δ Meteorology MAE (%)	Δ Soil MAE (%)	Δ Boundary MAE (%)	AC MAE (%)	Mean depth to GW (m)	Mean minimum anomalies (m)
df_0001	-32	11	9	-8	117	0.9	4.6	-0.8
df_0801	9	18	0	28	88	0.45	1.9	-0.4
df_1101	10	44	-1	11	232	0.92	11.1	-1.2
df_1201	19	14	50	18	325	0.93	12.3	-1.4
df_1601	12	-54	-41	-39	146	0.78	10.0	-0.5
df_1701	20	-30	72	26	186	0.86	3.6	-1.6
df_2101	2	-9	19	9	107	0.76	4.4	-0.9
df_2201	-6	-17	7	-17	126	0.84	4.4	-1.1
df_2601	0	25	-5	19	255	0.85	6.2	-0.7
df_2701	10	2	11	13	172	0.82	3.9	-2.4
df_2801	1	-9	-2	-2	57	0.59	2.8	-2.1
df_2901	5	16	8	-2	120	0.85	3.7	-1.3
df_3001	0	1	6	2	28	0.64	3.0	-1.7
df_3401	-4	0	3	-4	119	0.87	4.7	-1.0
df_04101	0	55	-3	15	80	0.81	6.3	-0.9
df_4101	46	43	-11	79	644	0.98	10.3	-1.3
df_4201	-7	-9	-30	-4	279	0.92	9.1	-1.4
df_4301	-11	-3	-14	-3	29	0.52	1.9	-0.5
df_4501	0	1	11	5	257	0.92	1.5	-2.4
df_4701	-2	2	12	3	24	0.47	1.2	-1.1
df_4901	-1	-1	3	-4	4	0.42	1.3	-0.9
df_5101	-3	-5	11	5	74	0.78	5.1	-0.7
df_5301	5	-4	10	2	425	0.96	5.8	-2.0
df_5601	0	-5	2	-5	47	0.48	1.6	-1.3
df_5701	-4	-21	34	4	231	0.95	6.0	-0.9
df_490001	-2	-3	-1	-3	33	0.67	4.4	-1.4

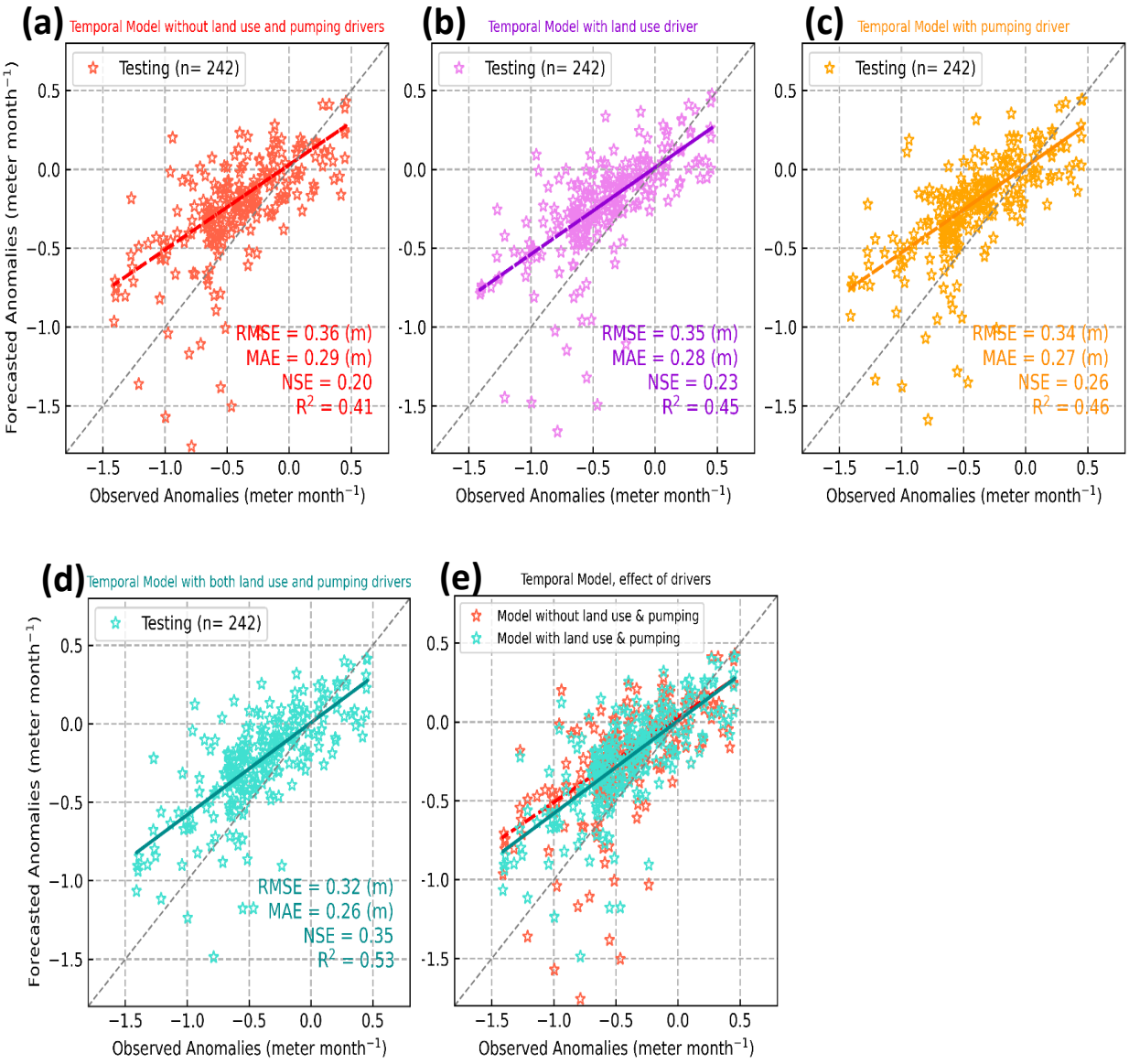

Figure. 4.11: Impact of drivers of land use and pumping is evaluated on the two-month forecast temporal anomalies model, calibration, or training time period: 2010-2017 (n=574), testing time period: 2016-2020 (n=242). Results are presented for the testing data. Evaluation metrics are shown for each case, (a) model without land use and irrigation pumping drivers, or baseline model, (b) model with land use information but without irrigation pumping data, (c) model with irrigation pumping data but without land use drivers, (d) model with both pumping and land use drivers, (e) Fig. 4.11a,b combined on one plot to show the effect of including or excluding drivers of land use and pumping.

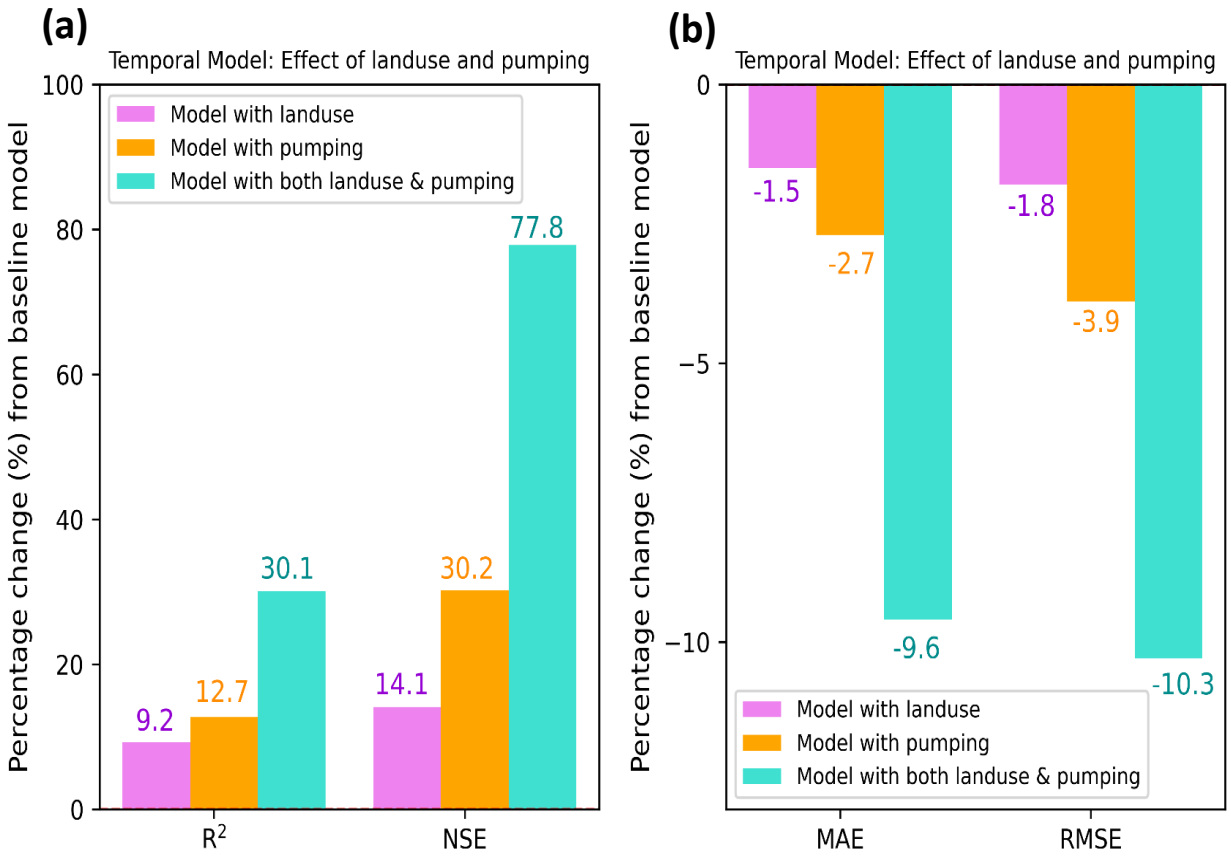

Figure 4.12: The temporal forecast model for two months out GW anomalies are compared to the baseline model to assess the influence of land use change and pumping (a) Percentage (%) change in R^2 and NSE for models with and without land use and pumping drivers, (b) Percentage (%) change in MAE and RMSE for models with and

4.4 Discussion

4.4.1 GW anomalies forecast models performance

Based on the performance measures for both the temporal (Fig. 4.5, Table 4.4) and spatial models (Fig. 4.8, Table 4.5), we find that machine learning based models can capture sub-seasonal variations in GW in heavily irrigated regions. These model sufficiently memorizes the dataset instead of generalizing it. MAE distributions across wells show that simulations differ from

observations due to complexities not accounted for in the model. However, no obvious spatial pattern clustering regions of high or low errors according to evaluation metrics like R^2 and NSE across the WCS. There would be an increase in MAE in those regions if one or more significant regional controls were missing from our model. We conclude that we are capturing the critical input parameters because the spatial distribution of error and R^2 values appears relatively uniform. According to our results, temporal models trained on subsets of data from all wells have broader distributions and lower mean R^2 and NSE than spatial models but narrower distributions and lower mean MAE and RMSE. In this case, the model benefited from the training values at each well. As a result, it is hardly surprising that the MAE is lower for the temporal model compared to the spatial model, which was calibrated based on only a few wells, and was reduced by its inability to represent heterogeneity.

Although the temporal model accurately captured GW anomalies in 2012 (Fig. 4.5a), the spatial model underestimated some extreme anomalies in 1965, 1974, 2002, and 2004. That effect might be caused by land use changes and high GW withdrawals that weren't accounted for in long-term forecast models. Groundwater level and extreme conditions were shown to be related by traditional hydrological models (Kidmose et al., 2013; Cai et al., 2021); however, machine learning models remain untested. According to our study, LSTM simulations performed better with a higher frequency of extreme conditions or variance, indicating that the model can learn how groundwater responds to external information (like baseflow recharging and precipitation).

Despite machine learning becoming popular in recent years for dealing with complex processes, little has been written about interpreting the predictions in hydrology (Campos-Taberner et al., 2020; Cheng et al., 2022). It has been shown that machine learning methods can explain complex processes in nature, including hydrology processes, according to studies on interpretation techniques (Montavon et al., 2018). For example, artificial neural networks have been used to explain, estimate and analyze crop yields in the Indian Wheat Belt (Wolanin et al., 2020), and to explain the relevant physical processes underlying seasonal precipitation (Gibson et al.,2021). Using different modeling scenarios, this study examines how catchment properties affect GW dynamics in the different WCS by excluding or including catchment-related drivers (Fig. 4.10, Fig. 4.12). In addition to providing insights into hydrology processes, the proposed interpretable machine learning approach also offers possibilities for hydrological catchment modeling in the future.

4.4.2 Model forecasts and hydrometeorological characteristics

Water levels in a catchment are affected by hydrometeorological characteristics, such as temperature, precipitation, and streamflow. A high temperature, for example, can increase ET, leading to a higher irrigation demand, and as a result, the groundwater level can drop faster (Hare et al., 2021). By excluding meteorological variables, MAE in some wells increased from 51-72%, as shown in this study (Fig. 4.10 d). Our study also observed a weak correlation between MAE for different wells and autocorrelation. Moreover, the correlation was reduced even further when meteorology-related variables were excluded. By influencing recharge, temperature and precipitation likely control groundwater levels in such cases. Moreover, precipitation and ET are

vital variables in quantifying temporal variations in GW, including those resulting from heavy rains and droughts.

Our study also revealed that by excluding SWE variables, MAE was increased as much as 46%, especially in wells at higher elevations. That is because cold areas are also affected by the snowmelt process, which can significantly impact the accuracy of hydrological models. Moreover, the snowmelt process can result in a lag effect on streamflow and groundwater recharge, reducing the accuracy of hydrological models (Carroll et al., 2018). Further, groundwater and streamflow recharge–discharge relationships in snow-domain areas are complicated (Harrison et al., 2021). Therefore, the lagged recharge caused by snowmelt and snowfall in catchments with lower mean temperatures and more snow, such as in New England, Mid-Atlantic, and Great Lakes, affect groundwater levels (Cai et al., 2021).

4.4.3 Relations between simulation results and catchment attributes

Groundwater anomalies and their spatiotemporal variability suggest that the geology, climate, and topography of catchment areas strongly influence groundwater flow differently at different sites. Despite the shorter delay time and faster response to external fluctuations for shallow wells, simulation results for the temporal model were better for deeper GW anomalies in our study. It could therefore be concluded that precipitation and baseflow infiltrated deeper subsurface groundwater more smoothly in this condition, which resulted in better simulation performance due to a higher groundwater recharge efficiency. Additionally, we found that meteorology-related variables, when excluded, significantly increased MAE in deep GW wells.

In traditional groundwater numerical models, the topography is often used as an important boundary condition (Bresciani et al., 2016). Therefore, it is important to understand how regional topography influences machine learning models. Condon & Maxwell (2015) found that groundwater flux is strongly influenced by topographic gradients in humid catchments with small topographic gradients. In another study by Cai et al., (2021), ML simulation results were more satisfactory in regions with lower mean elevations and slopes. Similarly, in our study, after removing topography and morphometry variables, two wells at higher elevations had the highest MAE increase (44-55%).

Regarding soil characteristics, which may affect groundwater recharging and soil water storage, a study showed that simulation performance was improved in regions with larger clay fractions and smaller sand fractions (CITATION?). By excluding soil-related and boundary-related variables from our study, we found that the wells with higher MAEs had a higher percentage of clay. Additionally, Šútor et al. (2010) found that regions with more uniform ratios of sand, silt, and clay had better simulation performance. Our study also found that excluding drivers from simulations led to fewer errors if the ratio of sand and clay was uniform. Indeed, the WCS aquifer is mainly sand, but simulation errors tend to increase when the clay percentage is too high or too low.

4.4.4 Land use change and irrigation pumping effect

Pumping or irrigation demand has the highest relative importance to GW anomalies forecast compared to land use change (Fig. 4.11, Fig. 4.12). There is a direct connection between pumping and groundwater level change based on simple physics, but initially, it wasn't clear

whether irrigation demand would have a stronger impact on groundwater anomalies than land use change since irrigation demand is determined by land use changes. Groundwater dynamics are closely related to agricultural irrigation in the WCS. The infiltration of irrigation water into the groundwater in this region is vital for replenishing the groundwater. It is mainly from the observation that groundwater levels drop during irrigation time but rise after irrigation. Furthermore, this study also found that drivers of land use change and irrigation pumping improved forecasts of shallow GW negative anomalies compared to deep GW negative anomalies. In WCS, where agricultural irrigation relies on shallow aquifers, shallow groundwater can subsidize water availability during droughts and increase crop yield.

4.5 Limitations and future directions

Unlike physical models, our modeling framework does not provide information about flux estimates and residence time calculations. Further, a new empirical model would be required to account for changes to the hydrologic system that were not included in our input parameters, such as managed aquifer recharge or imported water. In contrast, the physical model can be modified to include these additional sources and processes. As a result, each type of model excels in certain situations and applications. Furthermore, machine learning models need extensive training data to develop meaningful relationships between input and response variables. Providing data for machine learning models and mapping is a significant challenge. Data-deficient areas, particularly those with different hydroclimatic and geological settings, may not have adequate GW irrigation pumping or GW fluctuation data, hindering our methodology. Irrigation pumping data prior to 2010 would also have been helpful in our study.

It is important to note that ML methods depend highly on training samples to achieve their best performance. Because extreme anomalies values are bound to be limited samples, we observed that the overall MAE for the spatial model was higher than for the temporal model. There are likely to be errors in both long-term spatial and temporal models because irrigation data were unavailable before 2010 or because the model did not account for complexity, such as an increase in surface water capture during the study period as groundwater development increased rapidly in the WCS. Alternatively, model construction and parameterization uncertainties (e.g., boundary errors) could contribute to the error.

While physics-based models can be more general and interpretable (Zhao et al., 2019), they still have limitations in extracting optimal information from data. In contrast, ML models can take full advantage of available data but lack physical constraints (e.g., conservation of energy, diffusion laws) and interpretability (Parisouj et al., 2022). A most promising and challenging approach in hydrology appears to be the combination of physics-based models and machine learning. Additionally, the effectiveness of hybrid models, which combine physics-based and machine-learning models, in predicting GW dynamics remains to be determined. Nevertheless, a physics-constrained ML model can be developed to forecast GW dynamics in the future. One of the main advantages of such hybrid models is that they can predict GW anomalies and pure ML models with similar accuracy, yet hybrid model forecasting conserves water resources, respects boundary conditions and fluxes, and is more generalizable during extreme events. Such a method could be used to monitor droughts and heat waves globally.

4.5 Conclusion

We propose a recurrent neural network approach to forecast GW anomalies two months in advance. At 26 wells within the Wisconsin central sands (WCS), we evaluate and discuss the model's capability for predicting GW level changes from long-term averages. Pumping of groundwater in that region has caused groundwater levels to decline across large areas, causing concern among water resource managers. The region was modeled both temporally and spatially. A wide range of hydro-meteorological and geological variables are used in these models, along with piezometer data on GW fluctuations. Recording durations for these piezometers vary: some have a 50-year recording duration, while others are abandoned after being used for only 3 to 10 years. The time period of 1958-2020 was used for the development of the models. Human activities such as land use changes and irrigation water withdrawals were evaluated along with other catchment-related drivers. Training and testing data were used to assess the performance of spatial and temporal models.

Based on this study's use of LSTM for forecasting GW anomalies, the following conclusions are drawn:

- Bboth spatial (across wells) and temporal (one well at a time) models were able to predict GW anomalies with reasonable accuracy, with R^2 ~ 0.80 and MAE 0.07-0.34 m $month^{-1}$. The temporal model performed better than the spatial one in terms of overall accuracy. This model has a reasonable architecture, and the LSTM layer maintains previous knowledge and contributes to learning time series data.

- Based on the spatiotemporal variability of groundwater anomalies, there is a strong but variable influence of catchment-related drivers at different locations. For example, whiel boundary condition of the previous two-month GW were always an important driver, wells varied on relative important of drivers related to soil characteristics, meteorology, topography, and SWE.
- Irrigation demand played a more significant role in forecasting groundwater anomalies than land use change. As a result, it will be crucial to quantify and account for climate change and agricultural intensification when projecting future groundwater availability.

As a result of its integration of robust data processing by PCAs and input variable selection techniques with LSTM models, we were able to capture the impacts of the potential predictor variables on groundwater levels at multiple-well sites using our modeling framework. As a result of a changing climate and increased agricultural demand for water, these results may be helpful in identifying areas that are vulnerable to groundwater depletion. In agricultural regions where groundwater levels are abundant but subsurface parameters are largely unknown, our modeling framework is expected to be particularly useful enhancement over physical modeling methods. Groundwater level changes can be predicted using the forecasting model, especially in areas where physical hydrogeologic models are challenging to implement, providing a potential for season GW depletion risk analysis.

In this study, machine learning is demonstrated to have powerful nonlinear fitting capabilities. As a result, data-driven approaches will become more trusted, and more applications

in hydrological sciences will result from increased studies on machine learning and the general accuracy of model performance. However, machine learning models have been criticized the most for their black-box characteristics, which remain unchanged. In other words, although machine learning models, in theory, can provide accurate simulation results, the physical meaning of the models cannot be directly understood. Furthermore, groundwater levels fluctuate due to various factors, including melting snow and river discharge, as well as internal factors. For example, due to the lag effect of snowmelt, their overall simulation performances are not particularly favorable in regions where snow falls for a considerable proportion of precipitation,. Therefore, machine learning models incorporating physical constraints should be a future direction for groundwater dynamics simulations.

An alternative solution to the complex three-dimensional problems of groundwater levels in specific catchments is provided in this study. WCS aquifers are subject to a range of stresses, including groundwater withdrawal and land-use changes in recharge zones, and their hydrology geology is complex. Machine learning was used in this example to forecast GW anomalies for a complex basin, highlighting the importance of gathering reliable and long-term GW-level data. Machine learning algorithms will be an ideal tool to support decision-making in groundwater management as people worldwide become increasingly aware of the need for accurate and long-term groundwater monitoring. Overall, we demonstrate ML trained on long-term groundwater levels can be used as a measure of hydrologic and anthropogenic influence on GW levels and predict seasonal stresses to GW as a result of human decisions about water use for agriculture.

4.6 Acknowledgment

We acknowledge support from the Wisconsin Groundwater research and monitoring program, the Wisconsin Department of Natural Resources, and the UW Center for Climatic Research Climate, People, and Environment Program.

4.7 Bibliography

Abatzoglou, J.T., Williams, A.P., Boschetti, L., Zubkova, M., Kolden, C.A., 2018. Global patterns of interannual climate–fire relationships. Glob. Chang. Biol. 24. https://doi.org/10.1111/gcb.14405

Akbari Asanjan, A., Yang, T., Hsu, K., Sorooshian, S., Lin, J., Peng, Q., 2018. Short-Term Precipitation Forecast Based on the PERSIANN System and LSTM Recurrent Neural Networks. J. Geophys. Res. Atmos. 123. https://doi.org/10.1029/2018JD028375

Al-Yaari, A., Ducharne, A., Thiery, W., Cheruy, F., Lawrence, D., 2022. The Role of Irrigation Expansion on Historical Climate Change: Insights From CMIP6. Earth's Futur. 10. https://doi.org/10.1029/2022EF002859

Bengio, Y., Simard, P., Frasconi, P., 1994. Learning Long-Term Dependencies with Gradient Descent is Difficult. IEEE Trans. Neural Networks. https://doi.org/10.1109/72.279181

Bergen, K.J., Johnson, P.A., De Hoop, M. V., Beroza, G.C., 2019. Machine learning for data-driven discovery in solid Earth geoscience. Science (80-.). https://doi.org/10.1126/science.aau0323

Blender, R., Fraedrich, K., 2006. Long-term memory of the hydrological cycle and river runoffs in China in a high-resolution climate model. Int. J. Climatol. 26. https://doi.org/10.1002/joc.1325

Bradbury, K., Krause, J., Fienen, M., Kniffin, M., 2017. A groundwater flow model for the Little Plover River Basin in Wisconsin's Central Sands, Wisconsin Geological and Natural History Survey.

Bresciani, E., Gleeson, T., Goderniaux, P., de Dreuzy, J.R., Werner, A.D., Wörman, A., Zijl, W., Batelaan, O., 2016. Groundwater flow systems theory: research challenges beyond the

specified-head top boundary condition. Hydrogeol. J. 24. https://doi.org/10.1007/s10040-016-1397-8

Brodrick, P.G., Davies, A.B., Asner, G.P., 2019. Uncovering Ecological Patterns with Convolutional Neural Networks. Trends Ecol. Evol. https://doi.org/10.1016/j.tree.2019.03.006

Brown, J.F., Pervez, M.S., 2014. Merging remote sensing data and national agricultural statistics to model change in irrigated agriculture. Agric. Syst. 127. https://doi.org/10.1016/j.agsy.2014.01.004

Cai, H., Shi, H., Liu, S., Babovic, V., 2021. Impacts of regional characteristics on improving the accuracy of groundwater level prediction using machine learning: The case of central eastern continental United States. J. Hydrol. Reg. Stud. 37. https://doi.org/10.1016/j.ejrh.2021.100930

Campozano, L., Mendoza, D., Mosquera, G., Palacio-Baus, K., Célleri, R., Crespo, P., 2020. Wavelet analyses of neural networks based river discharge decomposition. Hydrol. Process. 34. https://doi.org/10.1002/hyp.13726

Carroll, R.W.H., Bearup, L.A., Brown, W., Dong, W., Bill, M., Willlams, K.H., 2018. Factors controlling seasonal groundwater and solute flux from snow-dominated basins. Hydrol. Process. 32. https://doi.org/10.1002/hyp.13151

Cheng, S., Cheng, L., Qin, S., Zhang, L., Liu, P., Liu, L., Xu, Z., Wang, Q., 2022. Improved Understanding of How Catchment Properties Control Hydrological Partitioning Through Machine Learning. Water Resour. Res. 58. https://doi.org/10.1029/2021WR031412

Clark, M.P., Fan, Y., Lawrence, D.M., Adam, J.C., Bolster, D., Gochis, D.J., Hooper, R.P., Kumar, M., Leung, L.R., Mackay, D.S., Maxwell, R.M., Shen, C., Swenson, S.C., Zeng, X., 2015. Improving the representation of hydrologic processes in Earth System Models. Water Resour. Res. 51. https://doi.org/10.1002/2015WR017096

Clark, M.P., Nijssen, B., Lundquist, J.D., Kavetski, D., Rupp, D.E., Woods, R.A., Freer, J.E., Gutmann, E.D., Wood, A.W., Brekke, L.D., Arnold, J.R., Gochis, D.J., Rasmussen, R.M., 2015. A unified approach for process-based hydrologic modeling: 1. Modeling concept. Water Resour. Res. 51. https://doi.org/10.1002/2015WR017198

Csillik, O., and Drăguţ, L. 2018. Towards a global geomorphometric atlas using Google Earth Engine. Geomorphometry. Available online at: http://2018.geomorphometry.org/Csilik_Dragut_2018_geomorphometry.pdf

Cuthbert, M.O., Gleeson, T., Moosdorf, N., Befus, K.M., Schneider, A., Hartmann, J., Lehner, B., 2019. Global patterns and dynamics of climate–groundwater interactions. Nat. Clim. Chang. https://doi.org/10.1038/s41558-018-0386-4

Deangelis, A., Dominguez, F., Fan, Y., Robock, A., Kustu, M.D., Robinson, D., 2010. Evidence of enhanced precipitation due to irrigation over the Great Plains of the United States. J. Geophys. Res. Atmos. 115. https://doi.org/10.1029/2010JD013892

Deines, J.M., Kendall, A.D., Hyndman, D.W., 2017. Annual Irrigation Dynamics in the U.S. Northern High Plains Derived from Landsat Satellite Data. Geophys. Res. Lett. 44. https://doi.org/10.1002/2017GL074071

Döll, P., Hoffmann-Dobrev, H., Portmann, F.T., Siebert, S., Eicker, A., Rodell, M., Strassberg, G., Scanlon, B.R., 2012. Impact of water withdrawals from groundwater and surface water on continental water storage variations. J. Geodyn. 59–60. https://doi.org/10.1016/j.jog.2011.05.001

Ekblad, L., Herman, J.D., 2021. Toward Data-Driven Generation and Evaluation of Model Structure for Integrated Representations of Human Behavior in Water Resources Systems. Water Resour. Res. 57. https://doi.org/10.1029/2020WR028148

Gibson, P.B., Chapman, W.E., Altinok, A., Delle Monache, L., DeFlorio, M.J., Waliser, D.E., 2021. Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. Commun. Earth Environ. 2. https://doi.org/10.1038/s43247-021-00225-4

Gleeson, T., Marklund, L., Smith, L., Manning, A.H., 2011. Classifying the water table at regional to continental scales. Geophys. Res. Lett. 38. https://doi.org/10.1029/2010GL046427

Graves, A., Mohamed, A.R., Hinton, G., 2013. Speech recognition with deep recurrent neural networks, in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. https://doi.org/10.1109/ICASSP.2013.6638947

Hare, D.K., Helton, A.M., Johnson, Z.C., Lane, J.W., Briggs, M.A., 2021. Continental-scale analysis of shallow and deep groundwater contributions to streams. Nat. Commun. 12. https://doi.org/10.1038/s41467-021-21651-0

Harrison, H.N., Hammond, J.C., Kampf, S., Kiewiet, L., 2021. On the hydrological difference between catchments above and below the intermittent-persistent snow transition. Hydrol. Process. 35. https://doi.org/10.1002/hyp.14411

Haucke, J., Clancy, K., Kraft, G., 2016. Tools to Estimate Groundwater Levels in the Presence of Changes of Precipitation and Pumping. J. Water Resour. Prot. 08. https://doi.org/10.4236/jwarp.2016.812084

Hengl, T., De Jesus, J.M., Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS One 12. https://doi.org/10.1371/journal.pone.0169748

Hintze, S., Glauser, G., Hunkeler, D., 2020. Influence of surface water – groundwater interactions on the spatial distribution of pesticide metabolites in groundwater. Sci. Total Environ. 733. https://doi.org/10.1016/j.scitotenv.2020.139109

Hu, X., Bürgmann, R., Xu, X., Fielding, E., Liu, Z., 2021. Machine-Learning Characterization of Tectonic, Hydrological and Anthropogenic Sources of Active Ground Deformation in California. J. Geophys. Res. Solid Earth 126. https://doi.org/10.1029/2021JB022373

Hurtt, G.C., Chini, L., Sahajpal, R., Frolking, S., Bodirsky, B.L., et al., 2020. Harmonization of global land use change and management for the period 850-2100 (LUH2) for CMIP6. Geosci. Model Dev. 13. https://doi.org/10.5194/gmd-13-5425-2020

Jiang, C., Guan, K., Pan, M., Ryu, Y., Peng, B., Wang, S., 2020. BESS-STAIR: A framework to estimate daily, 30m, and all-weather crop evapotranspiration using multi-source satellite data for the US Corn Belt. Hydrol. Earth Syst. Sci. 24. https://doi.org/10.5194/hess-24-1251-2020

Kalantar, B., Al-Najjar, H.A.H., Pradhan, B., Saeidi, V., Halin, A.A., Ueda, N., Naghibi, S.A., 2019. Optimized conditioning factors using machine learning techniques for groundwater potential mapping. Water (Switzerland) 11. https://doi.org/10.3390/w11091909

Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., Kumar, V., 2017. Theory-guided data science: A new paradigm for scientific discovery from data. IEEE Trans. Knowl. Data Eng. 29. https://doi.org/10.1109/TKDE.2017.2720168

Keene, A.A., and Mitchell, P.D. 2010. Economic impact of specialty crop production and processing in Wisconsin. Univ. of Wisconsin Ext., Madison.

Kidmose, J., Refsgaard, J.C., Troldborg, L., Seaby, L.P., Escrivà, M.M., 2013. Climate change impact on groundwater levels: Ensemble modelling of extreme values. Hydrol. Earth Syst. Sci. 17. https://doi.org/10.5194/hess-17-1619-2013

Kraft, G.J., and Mechenich, D.J. 2010. Groundwater pumping effects on groundwater levels, lake levels, and streamflows in the Wisconsin Central Sands. Center for Watershed Science and Education, Univ. of Wisconsin, Stevens Point.

Kraft, G.J., Clancy, K., Mechenich, D.J., Haucke, J., 2012. Irrigation Effects in the Northern Lake States: Wisconsin Central Sands Revisited. Ground Water 50. https://doi.org/10.1111/j.1745-6584.2011.00836.x

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., Hochreiter, S., Nearing, G.S., 2019. Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. Water Resour. Res. 55. https://doi.org/10.1029/2019WR026065

Krishna, B., Satyaji Rao, Y.R., Vijaya, T., 2008. Modelling groundwater levels in an urban coastal aquifer using artificial neural networks. Hydrol. Process. 22. https://doi.org/10.1002/hyp.6686

Leidner, A.K., Buchanan, G.M., 2018. Satellite Remote Sensing for Conservation Action: Case Studies from Aquatic and Terrestrial Ecosystems, Satellite Remote Sensing for Conservation Action: Case Studies from Aquatic and Terrestrial Ecosystems. https://doi.org/10.1017/9781108631129

Liu, S., Shi, H., 2019. A Recursive Approach to Long-Term Prediction of Monthly Precipitation Using Genetic Programming. Water Resour. Manag. 33. https://doi.org/10.1007/s11269-018-2169-0

Lu, Y., Harding, K., Kueppers, L., 2017. Irrigation effects on land-atmosphere coupling strength in the United States. J. Clim. 30. https://doi.org/10.1175/JCLI-D-15-0706.1

Luyssaert, S., Jammet, M., Stoy, P.C., Estel, S.,et al., 2014. Land management and land-cover change have impacts of similar magnitude on surface temperature. Nat. Clim. Chang. 4. https://doi.org/10.1038/nclimate2196

Majumdar, S., Smith, R., Butler, J.J., Lakshmi, V., 2020. Groundwater Withdrawal Prediction Using Integrated Multitemporal Remote Sensing Data Sets and Machine Learning. Water Resour. Res. 56. https://doi.org/10.1029/2020WR028059

Malekzadeh, M., Kardar, S., Saeb, K., Shabanlou, S., Taghavi, L., 2019. A Novel Approach for Prediction of Monthly Ground Water Level Using a Hybrid Wavelet and Non-Tuned Self-

Adaptive Machine Learning Model. Water Resour. Manag. 33. https://doi.org/10.1007/s11269-019-2193-8

MardanDoost, B., Brookfield, A.E., Feddema, J., Sturm, B., Kastens, J., Peterson, D., Bishop, C., 2019. Estimating irrigation demand with geospatial and in-situ data: Application to the high plains aquifer, Kansas, USA. Agric. Water Manag. 223. https://doi.org/10.1016/j.agwat.2019.06.010

Martínez-Santos, P., Renard, P., 2020. Mapping Groundwater Potential Through an Ensemble of Big Data Methods. Groundwater 58. https://doi.org/10.1111/gwat.12939

Milly, P.C.D., Wetherald, R.T., Dunne, K.A., Delworth, T.L., 2002. Increasing risk of great floods in a changing climate. Nature 415. https://doi.org/10.1038/415514a

Mohanty, B.P., Skaggs, T.H., 2001. Spatio-temporal evolution and time-stable characteristics of soil moisture within remote sensing footprints with varying soil, slope, and vegetation. Adv. Water Resour. 24. https://doi.org/10.1016/S0309-1708(01)00034-3

Motew, M.M., Kucharik, C.J., 2013. Climate-induced changes in biome distribution, NPP, and hydrology in the Upper Midwest U.S.: A case study for potential vegetation. J. Geophys. Res. Biogeosciences 118. https://doi.org/10.1002/jgrg.20025

Naghibi, S.A., Pourghasemi, H.R., Dixon, B., 2016. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environ. Monit. Assess. 188. https://doi.org/10.1007/s10661-015-5049-6

Nie, W., Kumar, S. V., Peters-Lidard, C.D., Zaitchik, B.F., Arsenault, K.R., Bindlish, R., Liu, P.W., 2022. Assimilation of Remotely Sensed Leaf Area Index Enhances the Estimation of Anthropogenic Irrigation Water Use. J. Adv. Model. Earth Syst. 14. https://doi.org/10.1029/2022MS003040

Nocco, M.A., Kraft, G.J., Loheide, S.P., Kucharik, C.J., 2018. Drivers of Potential Recharge from Irrigated Agroecosystems in the Wisconsin Central Sands. Vadose Zo. J. 17. https://doi.org/10.2136/vzj2017.01.0008

Nocco, M.A., Smail, R.A., Kucharik, C.J., 2019. Observation of irrigation-induced climate change in the Midwest United States. Glob. Chang. Biol. 25. https://doi.org/10.1111/gcb.14725

Olaya, V., 2009. Basic land-surface parameters, Developments in Soil Science. https://doi.org/10.1016/S0166-2481(08)00006-8

Parisouj, P., Mokari, E., Mohebzadeh, H., Goharnejad, H., Jun, C., Oh, J., Bateni, S.M., 2022. Physics-Informed Data-Driven Model for Predicting Streamflow: A Case Study of the Voshmgir Basin, Iran. Appl. Sci. 12. https://doi.org/10.3390/app12157464

Pokhrel, Y., Hanasaki, N., Koirala, S., Cho, J., Yeh, P.J.F., Kim, H., Kanae, S., Oki, T., 2012. Incorporating anthropogenic water regulation modules into a land surface model. J. Hydrometeorol. 13. https://doi.org/10.1175/JHM-D-11-013.1

Rybalkin, V., Sudarshan, C., Weis, C., Lappas, J., Wehn, N., Cheng, L., 2020. Efficient Hardware Architectures for 1D- and MD-LSTM Networks. J. Signal Process. Syst. 92. https://doi.org/10.1007/s11265-020-01554-x

Siebert, S., Kummu, M., Porkka, M., Döll, P., Ramankutty, N., Scanlon, B.R., 2015. A global data set of the extent of irrigated land from 1900 to 2005. Hydrol. Earth Syst. Sci. 19. https://doi.org/10.5194/hess-19-1521-2015

Smail, B. 2016. Irrigated land use statistics for Wisconsin. Wisconsin Dep. of Nat. Resour., Water Use Section, Madison.

Smidt, S.J., Haacker, E.M.K., Kendall, A.D., Deines, J.M., Pei, L., Cotterman, K.A., Li, H., Liu, X., Basso, B., Hyndman, D.W., 2016. Complex water management in modern agriculture: Trends in the water-energy-food nexus over the High Plains Aquifer. Sci. Total Environ. 566–567. https://doi.org/10.1016/j.scitotenv.2016.05.127

Sudheer, K.P., Gosain, A.K., Ramasastri, K.S., 2002. A data-driven algorithm for constructing artificial neural network rainfall-runoff models. Hydrol. Process. 16. https://doi.org/10.1002/hyp.554

Šútor, J., Štekauerová, V., Nagy, V., 2010. Comparison of the monitored and modeled soil water storage of the upper soil layer: the influence of soil properties and groundwater table level. J. Hydrol. Hydromechanics 58. https://doi.org/10.2478/v10098-010-0026-9

Tang, J., Li, Y., Ding, M., Liu, H., Yang, D., Wu, X., 2022. An Ionospheric TEC Forecasting Model Based on a CNN-LSTM-Attention Mechanism Neural Network. Remote Sens. 14. https://doi.org/10.3390/rs14102433

Troy, T.J., Kipgen, C., Pal, I., 2015. The impact of climate extremes and irrigation on US crop yields. Environ. Res. Lett. 10. https://doi.org/10.1088/1748-9326/10/5/054013

USGS, 2016. The Water Cycle: Water Storage in the Atmosphere [WWW Document]. USGS.

Van Houdt, G., Mosquera, C., Nápoles, G., 2020. A review on the long short-term memory model. Artif. Intell. Rev. 53. https://doi.org/10.1007/s10462-020-09838-1

Vergopolan, N., Chaney, N.W., Beck, H.E., Pan, M., Sheffield, J., Chan, S., Wood, E.F., 2020. Combining hyper-resolution land surface modeling with SMAP brightness temperatures to obtain 30-m soil moisture estimates. Remote Sens. Environ. 242. https://doi.org/10.1016/j.rse.2020.111740

Wada, Y., Bierkens, M.F.P., De Roo, A., Dirmeyer, P.A.,et al., 2017. Human-water interface in hydrological modelling: Current status and future directions. Hydrol. Earth Syst. Sci. 21. https://doi.org/10.5194/hess-21-4169-2017

Weeks, E.P., and Stangland, H. 1971. Effects of irrigation on streamflow in the Central Sand Plain of Wisconsin. USGS, Water Resources Division, Washington, DC.

Weeks, E.P., Ericson, D.W., and Holt, C.L.R. 1965. Hydrology of the Little Plover River basin, Portage County, Wisconsin and the effects of water resource development. U.S. Gov. Print. Office, Washington, DC.

Willmott, C.J., Rowe, C.M., Mintz, Y., 1985. Climatology of the terrestrial seasonal water cycle. J. Climatol. 5. https://doi.org/10.1002/joc.3370050602

Wisser, D., Frolking, S., Douglas, E.M., Fekete, B.M., Vörösmarty, C.J., Schumann, A.H., 2008. Global irrigation water demand: Variability and uncertainties arising from agricultural and climate data sets. Geophys. Res. Lett. 35. https://doi.org/10.1029/2008GL035296

Yoon, H., Jun, S.C., Hyun, Y., Bae, G.O., Lee, K.K., 2011. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. J. Hydrol. 396. https://doi.org/10.1016/j.jhydrol.2010.11.002

Zahura, F.T., Goodall, J.L., Sadler, J.M., Shen, Y., Morsy, M.M., Behl, M., 2020. Training Machine Learning Surrogate Models From a High-Fidelity Physics-Based Model: Application for Real-Time Street-Scale Flood Prediction in an Urban Coastal Community. Water Resour. Res. 56. https://doi.org/10.1029/2019WR027038

Zewdie, G.K., Valladares, C., Cohen, M.B., Lary, D.J., Ramani, D., Tsidu, G.M., 2021. Data-Driven Forecasting of Low-Latitude Ionospheric Total Electron Content Using the Random Forest and LSTM Machine Learning Methods. Sp. Weather 19. https://doi.org/10.1029/2020SW002639

Zhao, W.L., Gentine, P., Reichstein, M., Zhang, Y., Zhou, S., Wen, Y., Lin, C., Li, X., Qiu, G.Y., 2019. Physics-Constrained Machine Learning of Evapotranspiration. Geophys. Res. Lett. https://doi.org/10.1029/2019GL085291

Chapter 5

5. Conclusion

5.1 Summary & Outlook

This dissertation began with background and motivation to estimate ET and GW dynamics in chapter 1. Existing approaches to measuring ET and GW were described and their limitations were discussed. Literature was presented about data-driven or machine learning approaches to estimating ET and GW. Research questions and scope for the dissertation were presented. In chapter 2, the research improves the parameterization of atmospheric emissivity of a regional irrigation scheduling program for evapotranspiration estimation using longwave radiation, meteorological, and eddy covariance measurements. Furthermore, the performance of land surface model-based ET_a, reanalysis, and remote sensing products varies at different spatio-temporal scales. In the Wisconsin Central Sands Regions (WCS), irrigation agriculture plays a fundamental role in supporting regional agricultural production and rural community development. For irrigation scheduling, farmers often rely on the Wisconsin Irrigation and Scheduling Program (WISP) model, developed by soil scientists at UW-Madison and based on satellite data. Outcomes of chapter 2 evaluate sources of bias in the regional Wisconsin Irrigation and Scheduling Program (WISP) model and develop a correction using eddy covariance (EC) observations. ET_a observations were made for five years (2018–2022), using EC systems in irrigated potatoes fields in Wisconsin Central Sands (WCS) region.

Multi-source ET datasets should be employed in future studies to enhance confidence in remotely sensed ET-related water budgets. Chapter 3 evaluated the potentials of random forest (RF) and long short-term memory (LSTM) neural networks to estimate and forecast daily ET for corn, soybeans, and potatoes in diverse agricultural farms. The proposed modeling architecture provides a field-scale, locally calibrated tool for accurate prediction and short-term forecasting of daily ET in areas lacking in situ ET, metrological, and biophysical data. Furthermore, since there is no clear understanding of minimum required predictors for accurate estimates of ET, our models with different sets of predictors can help to understand the need for essential or minimum drivers for different crop fields on various soil textures in areas with scarce data.

To assess whether ML models can also be useful for GW modeling, research in Chapter 4 is devoted to the GW anomalies forecasting in the Central Sands region of WI for Sustainable Management. Wisconsin is humid, though GW shortages have been reported in several regions, including the WCS. The shortage of GW in the WCS is caused by many factors, including climate variability, fast expansion of irrigation, increasing high-capacity wells, and increasing population. To provide GW dynamics forecast two month in advance at high spatial and temporal resolutions across large spatial extents (e.g. WCS and across the state) for sustainable water resource, there is an urgent need to develop a computationally-efficient and easy-to-update framework with the capability of efficient incorporation of recently developed and regularly updated environmental datasets (e.g., remote sensing maps of evapotranspiration and soil moisture). Outcomes of chapter 4 include developing a data-driven machine learning (ML) framework to model and forecast GW dynamics in the WCS from 1958 to 2020. The results of this study will guide

improving GW forecasting in Wisconsin and best management practices on GW quantity and quality related to the construction of high-capacity wells, land cover change, crop selection, and irrigation management under a changing climate.

5.2 Improving parameterization of a regional irrigation scheduling program for evapotranspiration estimation with eddy covariance measurements

In irrigated agriculture, estimating crop evapotranspiration on a watershed scale has become increasingly important for water management. Chapter 2 presents a method for correcting ET models that fail to account for energy interaction. LW_{net} radiation is underestimated by the default WISP, which leads to large biases (both in magnitude and sign) in ET_a estimation. The improved parameterization greatly impacts surface energy budgets for energy-limited regions. Our goal was to demonstrate how to calibrate and reduce errors in local WISP ET_a models simultaneously. WISP model calibration was achieved by measuring a densely irrigated area of WCS. Using clear sky atmospheric emissivity ($\varepsilon_{a,clr}$) corrections, we estimated LW_{net} and ET_a accurately. Calibrating the model reduced the ET_a simulation error compared to the FLUXNET observation. For potatoes at our training sites, ET_a ranged from 0.5-5.7 mm d-1 (mean 2.9 mm d^{-1}). WISP overestimated ET_a by 1.15 mm d-1 and by 0.08 NSE. Despite this, after correction, the MAE of WISP ET_a and observed ET_a can be seen to be in good agreement with an NSE of 0.52 and a MAE of 0.72 mm d^{-1}. Additionally, model performance was mainly consistent for validation sites and crops.

ET and hydrologic models can be substantially improved by incorporating our findings into uncertainty studies. Aside from evaluating current ET and PET models, it would be beneficial to

incorporate energy corrections for latent fluxes into the model to explicitly account for input energy uncertainty. Our corrections to a popular longwave parameterization should be applied to other models that rely on this model too. When in situ meteorological and radiative measurements are available, flux measurements can be skillfully used to evaluate these models. Consequently, the methodology presented here can be applied to estimating latent heat fluxes for evaluating models or estimating annual ET for analyzing water budgets. In this study, remote sensing evapotranspiration estimates were combined with EC observations in order to inform water resources management and planning.

5.3 Evaluation of prediction and forecasting models for evapotranspiration of agricultural lands in the Midwest U.S.

Using a machine learning data-driven network, we have developed a new framework for estimating and forecasting ET and its uncertainty for corn, soybeans, and potatoes in agricultural areas in the Midwest. For building the model, biophysical and meteorological information was gathered from ground observations and satellite sensors. This model uses data sets widely used in ET prediction studies and hydrological models (SWAT and HSPF) as ancillary data. We calibrated the proposed model using 13 field-based eddy covariance ET time series in the Midwest for the period 2003-2019. For the period 2003-2019, the model was evaluated in seven independent locations.

According to the evaluation results of the predicted models based on ET measurements collected at seven different sites, these ML models are capable of estimating daily ET with a

ubRMSE of 0.67-0.92 mm, Willmott's skill score of 0.80-0.90, and simulating spatial heterogeneity of agricultural parameters and crop water use dynamics. Estimates from the prediction model were reliable and comparable to those from the mechanistic model from NLDAS. As a result of this study, we find common and important input predictors of EVI, solar zenith angle, incoming SW radiation, and CumGDD. Forecasting future ET was more dependent on vapor pressure. Both rainfed and irrigated crops can be modeled using the proposed model. The results of our study support the use of machine learning, especially random forest approaches, for prediction and short-term forecasting of ET in rainfed and irrigated crops, which can be applied to irrigation management and water cycle monitoring. Although further evaluation of additional predictors may be needed to expand this work to tropical or semi-arid regions, overall, the results indicate that a general regional ET model can be developed regardless of soil characteristics or climate. This model framework can be used to predict ET and forecast ET in order to help policymakers allocate water sustainably for irrigation and growers to identify areas of water stress.

5.4 Spatial and Temporal Forecasting of Groundwater anomalies

In the WCS regions, groundwater levels have declined across large areas due to pumping, causing concern among water resource managers. We modeled the region both temporally and spatially. GW anomalies can be predicted two months in advance using recurrent neural networks. We evaluated and discussed the model's ability to predict GW level changes from 26 wells within the Wisconsin Central Sands (WCS) based on long-term averages. These models include hydro-meteorological and geological variables and piezometer data on GW fluctuations.

Many piezometers have a 50-year recording duration, while others are abandoned after 3 to 10 years, requiring careful experimental design to incorporate these into model training and testing. Models were developed using a period of 1958-2020. As well as other catchment-related drivers, human activities like land use changes and irrigation water withdrawals were evaluated. The performance of spatial and temporal models was assessed using training and testing data. With R^2 = 0.80 and MAE of 0.07-0.34 m $month^{-1}$, this study found that both spatial and temporal models could predict GW anomalies reasonably well.

In terms of overall accuracy, the temporal model outperformed the spatial model. An LSTM layer contributes to learning time series data and maintains previous knowledge in this model. At different locations, groundwater anomalies are strongly influenced by catchment-related drivers, but their influence is variable. Another important driver was the boundary condition of the previous two-month GW, followed by soil characteristics, meteorology, topography, and SWE. Forecasting groundwater anomalies was more dependent on irrigation demand than land use change. In order to protect groundwater availability, it will be crucial to quantify and account for climate change and agricultural intensification. Using our model framework, we were able to detect the impacts of potential predictor variables on groundwater levels at multiple-well sites by integrating robust data processing by PCAs and input variable selection techniques. The results of this study may assist in identifying areas at risk of groundwater depletion due to climate change and increased agricultural water demand. Comparing our modeling framework to physical modeling methods, it should be instrumental in agricultural regions where groundwater levels are abundant but subsurface parameters are

unknown. Forecasting models can be used to predict groundwater level changes, especially in areas where physical hydrogeologic models are challenging to use.

5.5 Limitations & Future work

As part of our ET prediction and forecasting approach, soil evaporation and plant transpiration were not separated. In the early stages of growing seasons, when ET is mainly determined by soil evaporation, this leads to uncertainty in ET estimates. The optimization of WISP can be affected by errors in field measurements of ET. For example, the application of eddy covariance measurements has been well-recognized for energy balance closure issues (Twine et al., 2000; Baldocchi et al., 2016; Eichelmann et al., 2018). We might be unable to close the energy balance at our sites due to incomplete storage calculations in the WCS. Due to the varying footprint sizes of flux towers, the scale mismatch between WISP ET and tower measurements can introduce additional uncertainties. In future studies, a rigorous footprint approximation will also be required (Anderson et al., 2018; Wong et al., 2021). The flux footprint model (Kljun et al. 2015) can be implemented in the future to determine the weight and extent of pixel windows at flux tower sites. The estimates of energy and ET can also be uncertain when water use varies across a region, especially during the rainy season or when plants are growing and senescent. There are differences in the canopy structures, leaf morphologies, and water absorption capacities of potatoes, corn, and soybeans.

By partitioning water fluxes and redistributing energy, canopy processes play a crucial role in biophysical processes (Bonan, 2008; Fan et al., 2019). Although transpiration through leaf stomata plays a significant role in terrestrial water cycles, vegetation leaves and stems intercept

precipitation and partition it into water evaporation, runoff, and infiltration (Alton et al., 2009; Schlesinger & Jasechko, 2014). Therefore, monthly ET bias is reduced when canopy surface resistance is limited to realistic values. Peters et al. (2010) attributed species' differences in transpiration per unit canopy area to differences in foliage structure and growing season length. Also, irrigating crops, fertilizing them, shading their canopy, and killing vines can significantly affect their water loss (Sharma & Irmak, 2022). The current ET model should be improved to incorporate these key processes and better represent energy interactions. The Priestley-Taylor method has, however, the advantage of being easier to implement than other more complex and computationally expensive methods. It is essential to estimate LW radiation in most energy balance studies and the interaction between the land surface and the atmosphere, including surface ET. A better understanding of ET processes can be gained by using more accurate sensors and EC towers in energy-limited areas. The robustness of our optimization approach is enhanced by multiple field measurements across the WCS.

We propose a modeling framework without root zone water dynamics as one of its limitations in terms of parameters. Under non-irrigated conditions, actual ET under wet conditions is assumed to be equal to potential crop ET when soils have enough water stored. As a result of limited soil moisture storage in dry conditions, actual ET may be reduced, and plant ET is more dependent on soil moisture. Also, prior seven and 30 day precipitation were particularly important for non-irrigated crops as soil moisture proxy predictors for future ET. Non-irrigated sites may experience a more rapid loss of soil moisture during heat waves during dry years (e.g., 2012, 2010). Furthermore, ML models do not perform well when extrapolated to extreme or rare

conditions. We saw these results in Fig. 14 for Chapter 3 for extreme events in a dry year (2012) and a wet year (2017). With ongoing anthropogenic climate change, flash droughts and flooding may become more common but missed in predictions due to the tendency of ML models to "regress to the mean". Using MSE as the objective calibration function, Gupta et al, (2009) found this result is more likely. In contrast to physical models, our GW anomalies forecast model does not provide flux estimates or residence time calculations. As a result, a new empirical model would need to be developed that considers changes to the hydrological system that are not included in our entry parameters, such as managed aquifer recharge or imported water. In contrast, the physical model can be modified to include these additional sources and processes. Because of this, certain types of models are more suitable for certain situations and applications than others.

An ML model like RF or LSTM shows better generalization than linear models and can be used to compare an ANN or autoregressive model with a one-layer ML model (Fang et al., 2017). However, ML models have limitations when it comes to ET modeling. Models here, for instance, are calibrated locally. It is possible that some combinations of crops and soil types were not well trained, which could lead to inaccurate predictions of ET at those locations, despite the calibration pooling across multiple crop and soil types. A significant amount of training data is a limitation of ML models. It is possible to extract the influence of climatic cycles on ET by using long-term climatic data. In this way, models developed for domains with long-term flux tower locations are less sensitive to uncertainty and are more reliable in predicting ET. As long as the limitations and assumptions are valid, mechanistic models have an indisputable advantage of

estimating hydrological variables with limited training data. However, physics-based models still have limitations when it comes to extracting optimal information from data (Zhao et al., 2019). ML models, on the other hand, can make full use of available data but are not constrained by physical laws (e.g., conservation of energy, diffusion laws) and are difficult to interpret (Parisouj et al., 2022). Physics-based models combined with machine learning seem to be the most promising and challenging approach in hydrology. Furthermore, it is unknown whether hybrid models, which combine physics-based and machine-learning models, can accurately predict GW dynamics. Future GW dynamics may be better forecast using physics-constrained ML models and requires testing. It is important to note that to date such hybrid models are as accurate as pure ML models at predicting GW anomalies. Yet, since the models have other properties such as conserving water resources, respecting boundary conditions, and are more generalizable during extreme weather, they have merit for further consideration. Droughts and heat waves could be monitored globally with such a method.