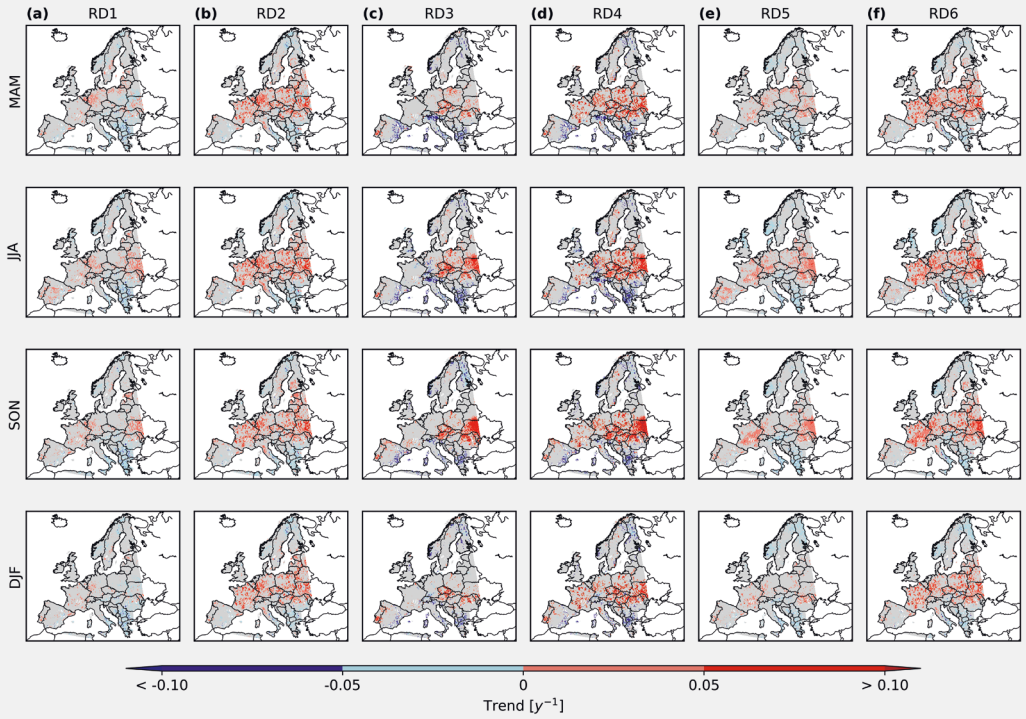

Machine learning for monitoring groundwater resources over Europe

Yueling Ma

Energie & Umwelt / Energy & Environment
Band / Volume 583
ISBN 978-3-95806-638-0

Mitglied der Helmholtz-Gemeinschaft

JÜLICH
Forschungszentrum

Schriften des Forschungszentrums Jülich
Reihe Energie & Umwelt / Energy & Environment Band / Volume 583

Forschungszentrum Jülich GmbH
Institut für Bio- und Geowissenschaften
Agrosphäre (IBG-3)

Machine learning for monitoring groundwater resources over Europe

Yueling Ma

Schriften des Forschungszentrums Jülich
Reihe Energie & Umwelt / Energy & Environment Band / Volume 583

ISSN 1866-1793 ISBN 978-3-95806-638-0

Bibliografische Information der Deutschen Nationalbibliothek.
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte Bibliografische Daten
sind im Internet über http://dnb.d-nb.de abrufbar.

Herausgeber und Vertrieb:	Forschungszentrum Jülich GmbH Zentralbibliothek, Verlag 52425 Jülich Tel.: +49 2461 61-5368 Fax: +49 2461 61-6103 zb-publikation@fz-juelich.de www.fz-juelich.de/zb
Umschlaggestaltung:	Grafische Medien, Forschungszentrum Jülich GmbH
Druck:	Grafische Medien, Forschungszentrum Jülich GmbH
Copyright:	Forschungszentrum Jülich 2022

Schriften des Forschungszentrums Jülich
Reihe Energie & Umwelt / Energy & Environment, Band / Volume 583

D 5 (Diss. Bonn, Univ., 2022)

ISSN 1866-1793
ISBN 978-3-95806-638-0

Vollständig frei verfügbar über das Publikationsportal des Forschungszentrums Jülich (JuSER)
unter www.fz-juelich.de/zb/openaccess.

Contents

Abstract

Groundwater (GW) is an important natural resource for Europe and the world, and has been affected by extreme weather and climate, e.g., summer heat waves and droughts, and human overexploitation. As climate change and human interventions increase, extreme events and GW depletion are expected to become more frequent and severe in many parts of Europe in the future, aggravating the vulnerability of GW systems. This emphasizes the necessity of GW monitoring in GW management. Up to date, however, it is still challenging to monitor GW at the large, continental scale, mainly due to the lack of water table depth (*wtd*) observations.

In order to address the challenge, the PhD work proposes an indirect, generic methodology based on advanced machine learning (ML) techniques, that are Long Short-Term Memory (LSTM) networks and transfer learning (TL), to produce reliable monthly *wtd* anomaly (wtd_a) estimates at the continental scale. The methodology is named LSTM-TL. While in this work, LSTM-TL has been implemented over Europe, it is transferable to other regions in the world. The methodology relies on the close connection between GW and other atmospheric and terrestrial compartments in the water cycle, using precipitation and soil moisture anomalies (pr_a and θ_a) as input, which have data available at large scales from, e.g., remotely sensed observations. Several steps were involved in the development of LSTM-TL for GW monitoring.

In the first step, LSTM networks were applied in combination with spatiotemporally continuous pr_a and wtd_a data from uncalibrated integrated hydrologic simulation results (named the TSMP-G2A data set) over Europe to capture the time-varying and time-lagged relationship between pr_a and wtd_a in order to obtain reliable networks to estimate wtd_a at the individual pixel level assuming that pr_a is a useful proxy for wtd_a. In most European regions, LSTM networks showed good skill with respect to the TSMP-G2A data set in predicting wtd_a with pr_a as input. The results indicated that the local factors, that are yearly averaged *wtd*, evapotranspiration (*ET*), soil moisture (θ), and snow water equivalent (*SWE*), had a significant impact on the performance of the LSTM networks. Moreover, the decrease in the network test performance at some pixels was attributed to a change in the temporal TSMP-G2A pr_a-wtd_a pattern during the study period.

In the second step, a number of input hydrometeorological variables, in addition to pr_a, were included in the construction of LSTM networks to arrive at improved wtd_a estimates at individual pixels over Europe in various experiments. All input and target data were derived from the TSMP-G2A data set. Improved LSTM networks were found with pr_a and θ_a as input. Considering θ_a strongly increased the network test performance particularly in the areas with $wtd \leq 3$ m (i.e., the major *wtd* category of Europe), suggesting the substantial contribution of θ_a to the estimation of wtd_a over Europe. The results highlight the importance to combine θ information with precipitation information in quantifying and predicting wtd_a.

In the final step, LSTM-TL was proposed for real-world applications. In LSTM-TL, LSTM networks were first trained on TSMP-G2A anomalies, and then, without additional training, utilized to estimate wtd_a with pr_a and θ_a from common observational datasets as input, thus, transferring knowledge from simulation results (i.e., the TSMP-G2A data set) to the observation-based estimation of wtd_a. Applying TL addressed the issue of scarce wtd_a observations ($wtd_{a,o}$) to train LSTM networks at the European scale. The implementation of LSTM-TL was based on two assumptions, that are i) the modeled relationship between wtd_a and input hydrometeorological variables (i.e., pr_a and θ_a) agreed well with the observed; and ii) the internal LSTM networks successfully captured the modeled relationship. The obtained wtd_a estimates were evaluated with collated *in-situ* wtd_a measurements from approximately 2,600 European GW monitoring wells, which demonstrated the good skill of LSTM-TL in estimating wtd_a. LSTM-TL was used for reconstructing monthly wtd_a from the early 1980s to the near present over Europe. The reconstructed wtd_a data exhibited seasonal wtd_a trends in different European regions in the past, contributing significantly to the understanding of historical GW dynamics at the continental scale over Europe, which has not been possible before.

The proposed LSTM-TL has three salient features. First, the methodology does not rely on $wtd_{a,o}$ to estimate wtd_a, which enables its usage over large regions even without *wtd* observations. Second, the methodology can be used to generate wtd_a estimates beyond the time period of the TSMP-G2A data set utilized for training, which is useful for reconstructing historical wtd_a and predicting future wtd_a at the continental scale. Third, once the internal LSTM networks are successfully trained, the methodology can be directly implemented without additional training, and thus, requires low computational cost in comparison to physically-based numerical simulation systems to generate new wtd_a estimates.

This PhD work presents a novel approach in the field of ML to estimate wtd_a in the absence of *wtd* observations, which advances significantly GW monitoring capacities at large scales. Since the TSMP-G2A data set provides a near-natural representation of the terrestrial water and energy cycles, the current implementation of LSTM-TL does not account for anthropogenetic impacts on GW dynamics. Nevertheless, LSTM-TL has been shown to produce reliable wtd_a estimates over Europe and can serve as an alternative methodology to *in-situ* wtd_a measurements. In addition to data reconstruction, the methodology can be employed for online GW monitoring and predictions, which is useful to GW management in Europe and beyond.

List of Symbols and Abbreviations

f_m	Modeled input-output relationship
f_o	Observed input-output relationship
f_{lstm}	The modeled input-output relationship captured by the LSTM networks in LSTM-TL
ET	Evapotranspiration
ET_a	Evapotranspiration anomaly
PFT	Plant functional type
pr_a	Precipitation anomaly
$pr_{a,m}$	Precipitation anomaly from modeling results
$pr_{a,o}$	Precipitation anomaly from observations
rs_a	River stage anomaly
S_t	Soil type
SWE	Snow water equivalent
SWE_{scaled}	Scaled yearly averaged snow water equivalent
wtd	Water table depth
wtd_a	Water table depth anomaly
$wtd_{a,lstm(m)}$	Water table depth anomaly estimates obtained by LSTM networks using modeling results as input
$wtd_{a,lstm(o)}$	Water table depth anomaly estimates obtained by LSTM networks trained on observations
$wtd_{a,lstm-tl}$	Water table depth anomaly estimates obtained by LSTM-TL
$wtd_{a,m}$	Water table depth anomaly from modeling results
$wtd_{a,o}$	Water table depth anomaly from observations
θ	Soil moisture
θ_a	Soil moisture anomaly
$\theta_{a,m}$	Soil moisture anomaly from modeling results
$\theta_{a,o}$	Soil moisture anomaly from observations

AI	Artificial intelligence
AL	Alps
ANN	Artificial neural network
BI	British Isles
CEC	Constant error carousel

CLM	Community Land Model
CORDEX	COordinated Regional Downscaling Experiment
COSMO	COnsortium for Small-scale MOdeling
CWT	Continuous wavelet transform
DL	Deep learning
EA	Eastern Europe
ECA&D	European Climate Assessment and Dataset
EEA	European Environment Agency
FFNN	Feedforward neural network
FR	France
GRACE	Gravity Recovery and Climate Experiment
GW	Groundwater
IP	Iberian Peninsula
JUWELS	Jülich Wizard for European Leadership Science
KGE	Kling-Gupta efficiency
LSTM network	Long Short-Term Memory network
MD	Mediterranean
ME	Mid-Europe
ML	Machine learning
MSE	Mean square error
NCAR	National Center for Atmospheric Research
NOAA	National Oceanic and Atmospheric Administration
NSE	Nash-Sutcliffe efficiency
OASIS3–MCT	Ocean Atmosphere Sea Ice Soil Model Coupling Toolkit
PRUDENCE	Prediction of Regional scenarios and Uncertainties for Defining EuropeaN Climate change risks and Effects
r	Pearson correlation coefficient
TL	Transfer learning
RMSE	Root mean square error
RNN	Recurrent neural network
SC	Scandinavia
SPEI	The Standardized Precipitation Evapotranspiration Index
SPI	The Standardized Precipitation Index
TSMP	Terrestrial Systems Modeling Platform
XWT	Cross-wavelet transform
α	Bias from the Pearson correlation coefficient

Chapter 1 Introduction

1.1 Background

Groundwater (GW) is an important natural resource for Europe and the world. About 75% of the residents in the European Union (European Commission, 2008) and billions of people worldwide (Alley et al., 2002) rely on GW for their drinking water supply. In addition, GW plays a crucial role in ensuring food security, adapting to climate variability, supporting biodiversity and sustaining surface water bodies (van der Gun, 2020). In recent years, however, global GW systems have been more or less affected as a consequence of extreme weather and climate, e.g., summer heat waves and droughts, and human overexploitation (Green et al., 2011; de Graaf et al., 2019; Jasechko and Perrone, 2021). As climate change and human interventions increase, the risks of extreme events and GW depletion are expected to escalate in many parts of Europe in the future (Guerreiro et al., 2018; Bloomfield et al., 2019; de Graaf et al., 2019), exacerbating the adverse impact on GW, thereby threatening environmental and socioeconomic sustainability and security. This emphasizes the need of effective and efficient GW monitoring over large regions.

Currently, continental-scale GW monitoring is lacking, mainly due to the scarcity of water table depth (*wtd*) observations, and fragmentation of data sets between different industries, and communal, state and federal agencies (Brauns et al., 2020). This inhibits the understanding of the connection of weather and climate extreme events with GW dynamics over large regions. As such, several alternative methodologies have been proposed to quantify and predict GW changes in the absence of *wtd* observations, using e.g., standardized meteorological drought indices over extended time scales (Bloomfield and Marchant, 2013; Kumar et al., 2016; Van Loon et al., 2017); terrestrial water storage anomalies derived from satellite observations (Rodell et al., 2018; Boergens et al., 2020); and physically-based numerical models (Tallaksen et al., 2009; Hartick et al., 2021). These methodologies can be or have been used to estimate GW anomalies at large scales, but they are questioned for their reliability, or require extensive computational resources (see Section 1.3 for more details).

As an alternative, machine learning (ML) techniques are able to automatically capture complex nonlinear relationship between GW and hydrometeorological variables (e.g., precipitation) from historical data, potentially resulting in more robust predictive models for GW anomalies. Relying on less background knowledge, ML techniques may be used to obtain accurate and reliable *wtd* anomaly (wtd_a) estimates comparable to physically-based numerical models, thereby significantly reducing the requirements for human involvement and expertise (Govindaraju, 2000). Moreover, they generally require less computational resources after successful training in routine applications. Malakar et al. (2021) provided a summary of previous studies, in which ML techniques were employed for GW resource modeling. These studies were mainly conducted at the aquifer and watershed scales. The application of ML to model GW at the large, continental scale is still not covered, which may be due to sparse GW observations available for training.

The present PhD work explores the potential of ML techniques to generate wtd_a estimates at the European scale based on different input hydrometeorological variables, and attempts to derive a ML-based methodology to obtain reliable estimation for wtd_a in space and time over Europe in the absence of *wtd* observations. It is worth noting that although the study area is Europe, the proposed methodology is transferable to other regions in the world. This chapter first presents the interactions between GW and other compartments in the water cycle, constituting the physical basis of the proposed methodology. Then, three common alternative methodologies for estimating GW anomalies are introduced, together with a discussion about their strengths and weaknesses. Next, the application of ML in GW resource modeling is reviewed. In what follows, the objective of the PhD work and the outline of each chapter are provided. Finally, the software tools and supercomputing environment adopted in the work are detailed.

1.2 Interactions between groundwater and other compartments in the water cycle

The water cycle (Figure 1.1) refers to the pathways of water through the Earth system including the atmosphere, land and oceans. Atmospheric water vapor condenses and falls over the land and oceans as precipitation. When precipitation arrives at the land surface, some of the water may flow over the surface in the form of surface flow, and some may infiltrate into the soil. A similar phenomenon occurs as snow melts. The surface flow collects locally into surface depression, or it runs directly into streams or rivers, which ends up into a larger water body, e.g., lakes or oceans. Except for retaining in the soil layers, the infiltrated water may travel rapidly through the near-surface soil layers into springs or neighboring streams or percolate slowly across the soil layers into GW. Vegetation takes up GW or soil water by means of root water uptake. In turn, evapotranspiration (*ET*) returns water to the atmosphere and closes the water cycle. *ET* encompasses direct evaporation from open water or the soil and land surface, transpiration of the biological water from plants, and sublimation from snow (Brutsaert, 2005).

As an important terrestrial compartment of the water cycle, GW is stored in the pore space of permeable geological units termed aquifers (Aeschbach-Hertig and Gleeson, 2012) and closely linked to other terrestrial and atmospheric compartments. Most aquifers are replenished by precipitation (diffuse recharge) and surface water (localized recharge). Diffuse recharge moves the precipitation falling on the land surface to the water table (i.e., the upper boundary of unconfined aquifers where the water pressure is atmospheric) through infiltration and percolation. Localized recharge transports water from surface water bodies to aquifers and is more heterogenous in space than diffuse recharge. In arid regions, diffuse recharge is often less pronounced (Alley et al., 2002). Lateral flow occurs between adjacent aquifers, which is also a potential source of GW recharge. In general, water not only moves from surface water bodies to aquifers (i.e., localized recharge), but also flows reversely as GW discharge, providing baseflows to rivers and lakes in dry periods (Fan et al., 2013). Hence, a two-way interaction exists between GW and surface water, influenced by factors such as the location of the surface water body relative to the aquifer, the characteristics of the river or lake bed and underlying materials, and the climate (Alley et al., 2002). In addition, shallow

GW sustains water in the soil and vegetations during dry periods, and thus, acts as a source term of *ET* to the atmosphere (York et al., 2002; Fan et al., 2013). Anthropogenic activities (e.g., GW pumping and irrigation) can also have a substantial impact on GW systems, leading to increased recharge and decreased discharge, which is outside the scope of this work.

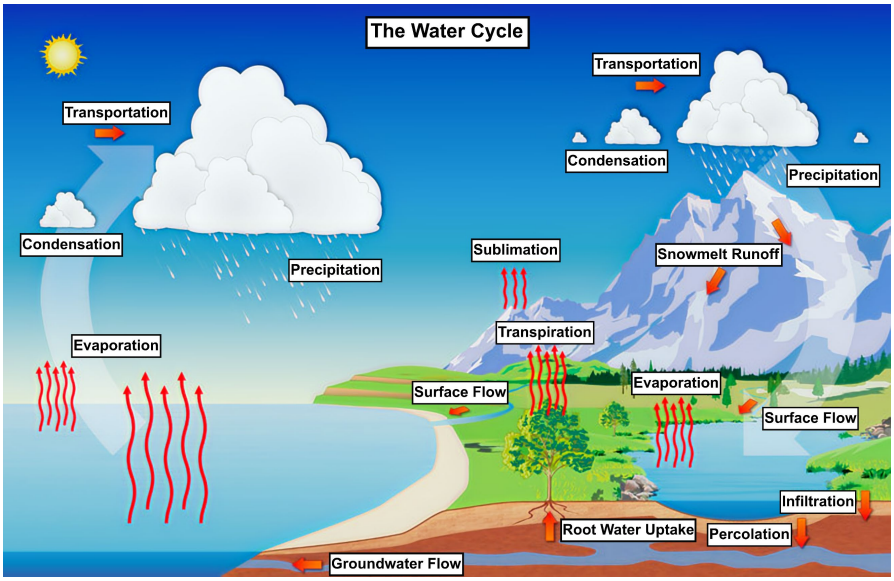

Figure 1.1: Sketch of the water cycle, adapted from the National Oceanic and Atmospheric Administration (NOAA) National Weather Service (2015).

Precipitation is a dominant source of GW, which provides diffuse recharge. Typically, only a small amount of precipitation becomes GW recharge, and the percentage of precipitation that recharges GW is controlled by factors including weather patterns, properties of surface soils, vegetation, local topography and the *wtd* (Alley et al., 2002). A time lag exists in the response of GW to precipitation changes, varying in space. The maximum dry/wet anomaly is attenuated in the movement of precipitation from the land surface to aquifers. As such, the land surface and soil layers behave as a low-pass filter to precipitation, resulting in smoothed and delayed GW anomaly signals (Van Loon, 2015). GW storage depends on antecedent storage conditions as well as recharge and discharge (Tallaksen et al., 2009; Van Loon, 2015). Precipitation deficits only affect GW recharge to some extent, and thus, may not be sufficient to induce GW drought.

1.3 Three common alternative methodologies for estimating groundwater anomalies

The lack of spatiotemporally continuous *wtd* observations is ubiquitous in GW studies. Several alternative methodologies have been proposed to estimate GW anomalies in the absence of *wtd* observations.

Because these methodologies do not rely on scarce *wtd* observations, they can be applied to large regions. Here, three commonly used methodologies are discussed briefly.

The first methodology is to use standardized meteorological drought indices over extended time scales to represent wtd_a, which consist of the Standardized Precipitation Index (SPI) introduced by McKee et al. (1993) and the Standardized Precipitation Evapotranspiration Index (SPEI) introduced by Vicente-Serrano et al. (2010). SPI and SPEI are obtained via the sum of corresponding monthly values (precipitation or precipitation minus potential *ET*) over different time scales (e.g., 3, 6, 9 and 12 months). The sum is fitted to a probability distribution, which is then transformed to a normal distribution. Thus, for the desired location and time scale, the mean and standard deviation of the obtained SPI or SPEI are zero and one, respectively (Mishra and Singh, 2010). This methodology assumes that the accumulation of meteorological drought signals (e.g., precipitation deficits) over a long period is representative for GW anomalies (Van Loon et al., 2017). It has two main advantages, which are i) easy to compute; ii) and comparable across different hydrometeorological regions. Hence, the methodology has been applied in many countries, e.g., Khan et al. (2008) in Australia, Fiorillo and Guadagno (2012) in Italy, ZAKHEM and KATTAA (2016) in Syria and Van Loon et al. (2017) in the Netherlands and Germany. However, some studies have shown that the methodology has difficulties in reproducing GW fluctuations due to the nonlinear translation of meteorological forcing such as precipitation to GW (Bloomfield and Marchant, 2013; Kumar et al., 2016; Uddameri et al., 2019). Moreover, as discussed in Section 1.2, simply considering meteorological information is not enough to quantify GW anomalies.

With the launch of the Gravity Recovery and Climate Experiment (GRACE) satellite mission in 2002, GRACE terrestrial water storage anomalies have become a common proxy for wtd_a, which show the variations in both near-surface water and GW storages. Chen et al. (2016) presented a review of GRACE applications in GW studies at the global scale. In addition, Thomas et al. (2017) proposed a new GW drought index - the GRACE Groundwater Drought Index, derived from GRACE terrestrial water storage anomalies. However, due to the coarse spatial resolution of the GRACE data (0.5°, about 55 km), the proxy may fail to represent spatially varying GW dynamics at the scale of small watersheds (Van Loon et al., 2017). Simulated near-surface water storage (including surface water, soil water and snow) is often used to isolate GW anomalies from GRACE terrestrial water storage anomalies (e.g., Rodell et al., 2007; Bonsor et al., 2018; Han et al., 2019), which introduces uncertainties in the simulation results for GW anomaly estimation. For instance, Van Loon et al. (2017) found that uncertainties in the simulation of soil moisture (θ) by various GLDAS models were the main reason for unreliable GRACE-based GW anomalies.

Physically-based numerical models constitute a tool for estimating GW anomalies. They are established based on physics equations to simulate processes in the terrestrial water and energy cycles. The first numerical models for modeling GW emerged around the 1960s, e.g., Tyson and Weber (1964) and Freeze and Witherspoon (1966). Later, many numerical models, such as MODFLOW (Hughes et al., 2017), FEFLOW (Diersch, 2014) and ParFlow (Maxwell et al., 2019), have been used to simulate GW flow

worldwide. Some of them have been coupled to land surface and atmospheric numerical models to obtain a more realistic representation of feedbacks between GW, the land surface and atmosphere, e.g., Sophocleous et al. (1999), Shrestha et al. (2014) and Simmer et al. (2015). Although physically-based numerical models can provide accurate estimates, they are time consuming and expensive to develop, apply and maintain, especially for high-resolution long-term simulations at large scales (Wunsch et al., 2018). The large amount of background knowledge and input data required for model setup is also a challenge.

1.4 Application of machine learning in groundwater resource modeling

ML is the study of computer algorithms that allow systems to extract knowledge from raw data without explicit programming (Goodfellow et al., 2017a; Müller and Guido, 2017a). ML belongs to the broad field of artificial intelligence (AI), which attempts to enable machines to mimic human behaviors. As a family of ML techniques, artificial neural networks (ANNs) are computing systems constituting many basic information-processing units (termed neurons) interconnected together (Haykin, 2009). Deep learning (DL) generally refers to large, multi-layer ANNs capable of learning and adapting directly from big raw data. The main difference between deep and non-deep neural networks is the depth of the network, often determined by the number of layers it contains (Shen, 2018; Malakar et al., 2021). The relationship between these different terminologies is shown in Figure 1.2.

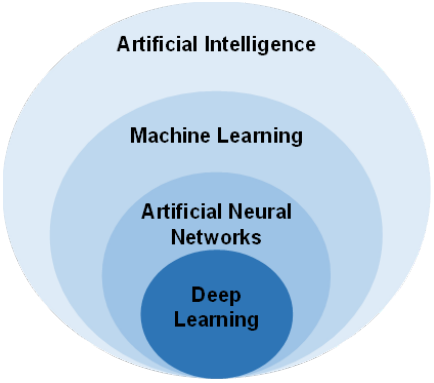

Figure 1.2: Relationship between AI, ML, ANNs and DL.

As described in Section 1.1, ML techniques can achieve comparable performance to physically-based numerical models but utilize much less background knowledge and computational resources. Benefiting from these advantages, they have been widely applied in GW resource modeling, such as predicting GW levels or terrestrial water storage. Most of these studies are devoted to simple ANNs especially feedforward neural networks (FFNNs) and their variants. Sun (2013) proposed FFNN ensembles to predict GW level changes one to three month ahead at multiple wells in the US based on precipitation, minimum and maximum temperatures and GRACE terrestrial water storage changes. They found that the FFNN

ensembles performed well at multi-month lead times by including GRACE terrestrial water storage changes as input. Gholami et al. (2015) successfully reconstructed GW level fluctuations from 1912 to 2013 at two wells in an alluvial aquifer in Iran based on tree rings and a FFNN. Yet, the method was only applicable to alluvial aquifers and non-arid regions. Sun et al. (2016) showed that a FFNN outperformed a multiple linear regression model in predicting GW levels up to seven days ahead in a tropical wetland in Singapore using surrounding reservoir levels and rainfall as input. Wunsch et al. (2018) presented the outstanding skill of nonlinear autoregressive networks with exogenous inputs (regarded as a combination of a FFNN and an autoregressive model with exogenous input) in short- and mid-term GW level predictions (one day up to six months ahead) at several wells in southwest Germany solely based on precipitation and temperature. More studies using shallow ANNs in GW resource prediction are referred to Malakar et al. (2021).

While shallow ANNs have shown great promise in modeling GW resources worldwide, they necessitate dedicated design for input data, i.e., manually defining the time lag in the response of target GW variables to input hydrometeorological forcings. The time lag generally varies in space, which increases the difficulties to apply shallow ANNs at large scales. In recent years, deep neural networks especially Long Short-Term Memory (LSTM) networks have attracted more and more attention in predicting GW levels, e.g., J. Zhang et al. (2018), Jeong et al. (2020), Supreetha et al. (2020), Müller et al. (2021), Wunsch et al. (2021) and Vu et al. (2021). Due to their internal memory, LSTM networks can automatically detect the long-short-term dependencies between input and target time sequences without explicitly defining the time lag. There are some debates about the superiority of LSTM networks over shallow ANNs for time sequence processing when training data sizes are small, such as Müller et al. (2021) and Wunsch et al. (2021). Nevertheless, simplified data preprocessing is still a big advantage for applying LSTM networks over large regions.

The above ML methods used for modeling GW resources, i.e., FFNNs and LSTM networks, are transferable to the estimation of wtd_a, which is the focus of this PhD work. Up to date, however, it is challenging to apply ML techniques to model GW at the large, continental scale, mainly due to sparse GW observations. Modeling GW resources based on ML is a supervised learning task that requires target data as a supplementary teacher signal to guide the training process. Therefore, even though GW data may not be utilized as input, they are still needed to train ML techniques, which is impossible for regions without GW observations.

1.5 Objective and outline

The objective of the PhD work is to develop a ML-based indirect methodology to produce reliable wtd_a estimates at the European scale in the absence of *wtd* observations in order to facilitate the pan-European GW monitoring. The inputs are chosen from hydrometeorological variables that have spatiotemporally continuous coverage over Europe. The methodology development relies on uncalibrated integrated hydrologic simulation results (named the TSMP-G2A data set) under a near-natural climatology of the

physical states of the terrestrial system, thereby neglecting the influences of human interventions on GW dynamics.

The PhD work is structured as follows.

Chapters 2 to 4 detail the development and application of the methodology (illustrated in Figure 1.3).

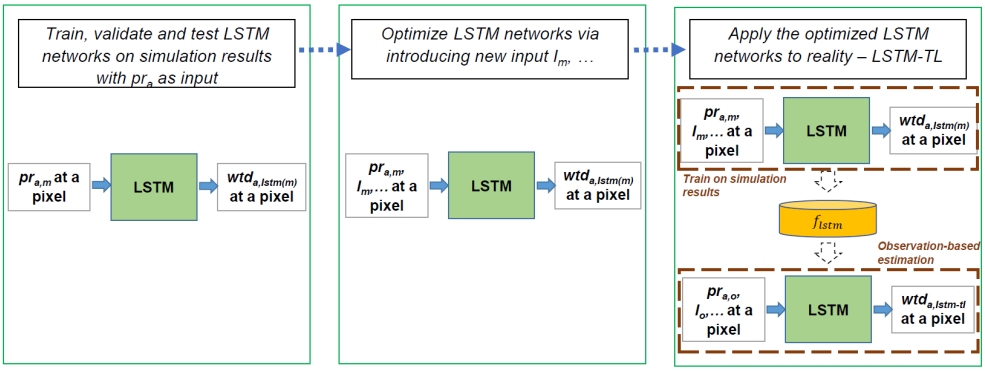

Figure 1.3: Schematic of the methodology development. The I indicates input variable. The variables with the subscripts m and o represent data from modeling results (also called simulation results) and observations, respectively. The $wtd_{a,lstm(m)}$ is the wtd_a estimate obtained by LSTM networks with simulation results as input. The $wtd_{a,lstm-tl}$ is the wtd_a estimate obtained by the proposed methodology LSTM-TL. The f_{lstm} is the input-output relationship learned from simulation results by the LSTM networks involved in LSTM-TL.

In Chapter 2, LSTM networks were employed in combination with spatiotemporally continuous pr_a and wtd_a data from the TSMP-G2A data set over Europe to capture the time-varying and time-lagged relationship between pr_a and wtd_a in order to obtain reliable networks to estimate wtd_a at the individual pixel level. The pr_a is the most common proxy of wtd_a, that is, using standardized meteorological drought indices over extended time scales to represent wtd_a.

In Chapter 3, a number of input hydrometeorological variables, in addition to pr_a, were investigated for the construction of LSTM networks to arrive at improved wtd_a estimates at individual pixels over Europe in various experiments. All input and target data were obtained from the TSMP-G2A data set.

Both Chapters 2 and 3 proposed and evaluated methodologies of modeled wtd_a ($wtd_{a,m}$) based on LSTM networks. In Chapter 4, the optimized LSTM networks from Chapter 3 were applied to real-world observations. Transfer learning (TL) was performed to overcome the challenge of scarce wtd_a observations ($wtd_{a,o}$) to train the LSTM networks over Europe. The proposed methodology, LSTM-TL, combined LSTM networks and TL to estimate wtd_a at the European scale in the absence of $wtd_{a,o}$ through knowledge transfer from the TSMP-G2A data set to the observation-based estimation. Two assumptions were made, that are i) the modeled relationship between wtd_a and input hydrometeorological variables was in good agreement

with the observed; and ii) the internal LSTM networks successfully captured the modeled relationship during training.

Finally, Chapter 5 summarizes the findings in the development of LSTM-TL, discusses the strengths and limitations of implementing ML techniques especially LSTM networks and LSTM-TL in hydrological studies, and provides conclusions and an outlook for future research. A statement about where to find the data and codes utilized or generated in this work is also provided.

1.6 Software tools and supercomputing environment

All the ML-based estimation models reported in this PhD work were developed using PyTorch (https://pytorch.org/), an open-source ML python library. The models were constructed on the supercomputer Jülich Wizard for European Leadership Science (JUWELS) at Forschungszentrum Jülich, including training, validation and testing. JUWELS is a modular supercomputer composed of a cluster module and a booster module. Each standard compute node on the cluster module utilized in this work has two Intel Xeon 24-core AVX512 CPUs with 96 GB of memory, and the compute nodes are interconnected with a Mellanox EDR InfiniBand (JSC, 2018). Training the ML-based models at all pixels with valid data (not including NaN or infinite values) over Europe cost ~ 150,000 core hours (~ 4 core hours at the individual pixel level). After successful training, the models required much less computational resources for validation and testing, i.e., ~ 160 core hours at the European scale. The obtained wtd_a estimates may be impacted by the versions of the applied python libraries and the platform on which the models were built.

Chapter 2 Using Long Short-Term Memory networks to connect water table depth anomalies to precipitation anomalies over Europe

The content of this chapter is adapted from:
Ma, Y., Montzka, C., Bayat, B. and Kollet, S.: Using Long Short-Term Memory networks to connect water table depth anomalies to precipitation anomalies over Europe, Hydrol. Earth Syst. Sci., 25(6), 3555–3575, doi:10.5194/hess-25-3555-2021, 2021.

Abstract

Many European countries rely on groundwater (GW) for public and industrial water supply. Due to a scarcity of near-real-time water table depth (*wtd*) observations, establishing a spatially consistent GW monitoring system at the continental scale is a challenge. Hence, it is necessary to develop alternative methods for estimating *wtd* anomalies (wtd_a) using other hydrometeorological observations routinely available near real time. In this work, we explore the potential of Long Short-Term Memory (LSTM) networks for producing monthly wtd_a using monthly precipitation anomalies (pr_a) as input. LSTM networks are a special category of artificial neural networks (ANNs) that are useful for detecting a long-term dependency within sequences, in our case time series, which is expected in the relationship between pr_a and wtd_a. In the proposed methodology, spatiotemporally continuous data were obtained from daily terrestrial simulations of the Terrestrial Systems Modeling Platform (TSMP) over Europe (hereafter termed the TSMP-G2A data set) with a spatial resolution of 0.11°, ranging from the year 1996 to 2016. The data were separated into a training set (1996-2012), a validation set (2013-2014), and a test set (2015-2016) to establish local networks at selected pixels across Europe. The wtd_a maps obtained by LSTM networks agreed well with TSMP-G2A wtd_a maps on spatially distributed dry and wet events, with 2003 and 2015 constituting drought years over Europe. Moreover, we categorized the test performances of the networks based on intervals of yearly averaged *wtd*, evapotranspiration (*ET*), soil moisture (θ), snow water equivalent (*SWE*), soil type (S_t), and dominant plant functional type (*PFT*). Superior test performance was found at the pixels with *wtd* < 3 m, *ET* > 200 mm, θ > 0.15 m^3m^{-3} and *SWE* < 10 mm, revealing a significant impact of the local factors on the ability of the networks to process information. Furthermore, results of the cross-wavelet transform (XWT) showed a change in the temporal pattern between TSMP-G2A pr_a and wtd_a at some selected pixels, which can be a reason for undesired network behavior. Our results demonstrate that LSTM networks are useful for producing high-quality wtd_a based on other hydrometeorological data measured and predicted at large scales, such as pr_a. This contribution may facilitate the establishment of an effective GW monitoring system over Europe that is relevant to water management.

2.1 Introduction

Groundwater (GW) is an essential natural resource, accounting for about 30% of the fresh water on Earth (Perlman, 2013) and sustains various domestic, agricultural, industrial and environmental uses, due to its widespread availability and limited vulnerability to pollution (Naghibi et al., 2016; Tian et al., 2016). According to the report of the European Environment Agency (EEA) in 1999, GW comprises over 50% of the public water supply in most European countries (EEA, 1999). GW systems are dynamic and adapt continuously to natural and anthropogenic stresses (Kenda et al., 2018). However, they have been affected in recent years as a consequence of frequent extreme weather conditions, e.g., severe droughts, and human overexploitation. Thus, effective and efficient GW management, especially under drought conditions, is required at the European scale to maintain environmental and socioeconomic sustainability.

Drought is characterized as the costliest natural hazard worldwide, resulting in significant societal, economic, and ecological impacts (Wilhite, 2000). The report of the EEA in 2016 demonstrated that drought had become a recurrent feature of the European climate; more droughts have occurred in some European countries than in the past, and their severity has also increased (EEA, 2016). Recent severe heat wave events in Europe occurred in 2003, 2015, and 2018, which led to several drought events covering most of the European continent (Norris, 2018). GW drought is a specific type of drought, impacting several important drought-sensitive sectors such as drinking water supply and irrigation (Van Loon et al., 2017). Hence, GW monitoring is ultimately indispensable over the European continent.

Effective GW monitoring requires accurate information on GW dynamics in space and time. One crucial variable for characterizing GW dynamics is the water table depth anomaly (wtd_a), reflecting anomalies in GW storage (Zhao et al., 2020), which is a key variable in GW drought analysis. The wtd_a is the deviation of the *wtd* value from the climatological average for a specified time period normalized by the climatological standard deviation, and can serve as a measure of GW drought. Commonly, *wtd* observations are measured *in-situ* in observation wells. However, to date, it is still a challenge to obtain near-real-time spatially continuous *wtd* observations over Europe (Van Loon et al., 2017; Bloomfield et al., 2018), and available data sets often suffer from uncertainties originating from unknown well bore and well installation specifics. Therefore, an alternative (indirect) method is needed to produce reliable area-wide wtd_a information over Europe.

Indirect methods rely on measurements of one (or more) hydrometeorological variable related to *wtd* via physical processes in the water cycle, such as infiltration and percolation. The precipitation anomaly (pr_a) is the most common variable used to model wtd_a, for which the calculation method is the same as wtd_a but based on precipitation. Precipitation is connected with GW via the process of percolation through soil layers. Thus, depending on evapotranspiration (*ET*) and the thickness of the vadose zone, a lag exists in the response of GW to precipitation. A considerable number of studies linked the accumulation of pr_a over extended timescales (e.g., 6 or 12 months) to wtd_a, often applying the Standardized Precipitation Index

(SPI) and the Standardized Precipitation Evapotranspiration Index (SPEI) to represent wtd_a (e.g., SPI: McKee et al., 1993; Thomas et al., 2015; SPEI: Vicente-Serrano et al., 2010; Van Loon et al., 2017). In these studies, equal weighting was assigned to the meteorological input in the derivation of the drought indices.

As an alternative, artificial neural networks (ANNs) are able to account for non-uniformly weighted, temporally lagged contributions of pr_a to wtd_a, potentially providing more robust prediction models. ANNs are one of the most widely used machine learning methods that have been inspired by biological neural systems, with many interconnected information processing units (i.e., neurons) (Haykin, 2009; Ma et al., 2019). ANNs adapt learnable parameters (i.e., weights and biases) in the links between neurons to achieve an appropriate input-output mapping based on observed data; this is also for complex nonlinear relationships. ANNs are not easily affected by input noise and are able to readjust their parameters when new information is included. More importantly, compared to physically-based numerical models, they necessitate little background knowledge, reducing the requirements for human involvement and expertise, thus enabling rapid hypothesis testing (Govindaraju, 2000; Shen, 2018; Sun and Scanlon, 2019).

Feedforward neural networks (FFNNs, also termed multilayer perceptrons in some of the literature) and their variants are commonly used ANNs for GW level modeling in previous studies (e.g., Yang et al., 1997; Nayak et al., 2006; Adamowski and Chan, 2011; Yoon et al., 2011; Mohanty et al., 2015; Gong et al., 2016; Sun et al., 2016). One major drawback of FFNNs is that they cannot preserve previous information, resulting in inefficiencies in handling sequential data (J. Zhang et al., 2018; Supreetha et al., 2020). To leverage the performance of FFNNs, the delay time in the network response needs to be estimated in advance.

Recurrent neural networks (RNNs) are a special type of ANN mainly designed for sequential data analysis. Through loops in their hidden layers, the information generated in the past flows back to neurons as the input of new computing processes (Karim and Rivera, 1992). Due to the ability to store information traveling through, RNNs can avoid the aforementioned preprocessing step of FFNNs and can, thereby, solve sequential data problems more efficiently. However, standard RNNs suffer from the exploding and vanishing gradient issues and often fail to exploit long-term dependencies between sequences, which is expected in the response of wtd_a to pr_a. These issues can be overcome by a variant of standard RNNs named Long Short-Term Memory (LSTM) networks (Supreetha et al., 2020). Although RNNs have been employed extensively in other science fields, particularly in natural language processing (D. Zhang et al., 2018), their application in hydrology is still in its infancy and has only recently received increasing attention (e.g., Kratzert et al., 2018; Shen, 2018; J. Zhang et al., 2018; Le et al., 2019; Sahoo et al., 2019). Thus, limited studies have been conducted to estimate GW fluctuations using RNNs, especially with LSTM networks.

The consistency of the temporal pattern between input and target variables is a prerequisite for the good performance of ANNs, including LSTM networks. Cross-wavelet transform (XWT) is a useful tool for visualizing the pattern changes between input and target variables, aiming to extract similarities of two time

series in time and frequency. The technique has been applied for time-frequency analysis in many publications (e.g., Adamowski, 2008; Prokoph and El Bilali, 2008; Banerjee and Mitra, 2014).

In this study, we utilized spatiotemporally continuous pr_a and wtd_a from integrated hydrologic simulation results of the Terrestrial Systems Modeling Platform (TSMP) over Europe (hereafter termed the TSMP-G2A data set and introduced in Section 2.2.4) in combination with LSTM networks to capture the time-varying and time-lagged relationship between pr_a and wtd_a in order to obtain reliable prediction models at the individual pixel level. The impact of local factors on the network behavior was also investigated, and the local factors studied were yearly averaged *wtd*, *ET*, soil moisture (θ), snow water equivalent (*SWE*), and soil type (S_t) and dominant plant functional type (*PFT*). In addition, we implemented XWT on both TSMP-G2A pr_a and wtd_a series for time-frequency analysis to gain insight into the internal characteristics of the obtained networks.

This chapter is organized as follows: in Section 2.2 (Methodology), we first present a conceptual model of GW balance to theoretically derive the relationship between pr_a and wtd_a and then briefly introduce the architecture of the proposed LSTM networks and continuous and cross-wavelet transform. This is then followed by detailed information on our study area and data set as well as a generic workflow to construct local LSTM networks at selected pixels over Europe. Section 2.3 (Results and discussion) shows reproduced wtd_a maps for GW drought analysis, discusses the impact of local factors on the network behaviors and investigates the network performances at the local scale, before completing the chapter with Section 2.4 (Summary and conclusions).

2.2 Methodology

LSTM networks were applied to estimate monthly wtd_a over the European continent, using monthly pr_a as input. We constructed the networks at the individual pixels and analyzed temporal patterns between TSMP-G2A pr_a and wtd_a using XWT. In this section, we briefly recall the conceptual model of GW balance, introduce the principle of LSTM networks and the application of XWT, and describe the study area and data set before presenting a universal workflow to establish the proposed LSTM networks locally at selected pixels.

2.2.1 Conceptual model of groundwater balance

The subsurface water balance can be described by a control volume that contains the vadose zone and an unconfined aquifer that is closed at the bottom (Figure 2.1). Note that areas with surface water are not taken into account in this study, and the impact of anthropogenic activities such as GW abstraction is neglected. Flows in and out of the control volume are precipitation and *ET* across the land surface and lateral flows in the subsurface. These flows are balanced by changes in the water stored in the vadose zone and the unconfined aquifer.

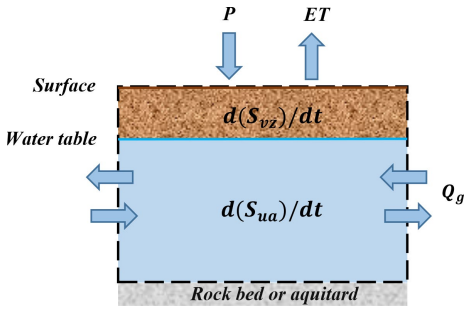

Figure 2.1: Conceptual model of GW balance over a control volume (after Maxwell, 2010). P is precipitation, ET is actual *ET*, Q_g is the lateral GW flow, S_{vz} and S_{ua} are the water storages in the vadose zone and the unconfined aquifer, respectively, and t is time.

The GW balance equation for the conceptual model is given in Eq. (2.1) as follows:

$$d(S_{vz})/dt + d(S_{ua})/dt = P - ET + \boldsymbol{Q_g} \ . \tag{2.1}$$

Rearranging Eq. (2.1), will result in Eq. (2.2), as follows:

$$d(S_{ua})/dt = P - ET + \boldsymbol{Q_g} - d(S_{vz})/dt \ , \tag{2.2}$$

where, P is precipitation [millimeters per month], and ET is actual *ET* [millimeters per month], and $\boldsymbol{Q_g}$ is the lateral GW flow [millimeters per month]. S_{vz} and S_{ua} are the water storages in the vadose zone [millimeters] and the unconfined aquifer [millimeters], respectively, and t is time [months].

The term on the left-hand side and the first term on the right-hand side in Eq. (2.2) indicate an explicit relationship between the fluctuation of S_{ua} and precipitation, providing the theoretical basis of this study. In the case of large continental watersheds (i.e., $\boldsymbol{Q_g} = 0$), the difference between precipitation and *ET* is equal to the total variations in S_{vz} and S_{ua}. Note that we explicitly separated the water storage term of the vadose zone from the unconfined aquifer to highlight the transient impact of unsaturated storage on the relationship between $d(S_{ua})/dt$ and (precipitation - *ET*).

2.2.2 Long Short-Term Memory networks

In this study, we employed LSTM networks with the same architecture of hidden neurons as Gers et al. (2000), which is shown in Figure 2.2. As a category of RNNs, LSTM networks have loops in their hidden layers that facilitate hidden neurons to weigh not only new inputs but also earlier outputs internally for predictions. Hence, similar to other RNNs, they are considered to have memory. Compared with standard RNNs, LSTM networks add a constant error carousel (CEC) and three gates that are the input, forget and output gates in their hidden neurons (see Figure 2.2), in order to overcome the exploding and vanishing gradient issues. For a detailed description of the functions of these components, the reader is referred to Hochreiter and Schmidhuber (1997) and Gers et al. (2000). Benefiting from the interaction of these

components, LSTM networks show great promise for studying long-term relationships between time series. They have the ability to capture dependencies over 1000 time steps, outperforming standard RNNs whose upper boundary of reliable performances is only 10 time steps (Hochreiter and Schmidhuber, 1997; Kratzert et al., 2018). The response of wtd_a to pr_a is expected to exhibit a long time lag, especially in case of deep aquifers, and thus LSTM networks are an appropriate type of network to use here. In addition, compared with traditional physically-based numerical models, the proposed LSTM networks require less computational time and background knowledge to perform the simulations. Moreover, when the proposed LSTM networks are available, we only need the pr_a data to estimate wtd_a, which are available from bias-corrected operational forecasts and reanalysis data sets.

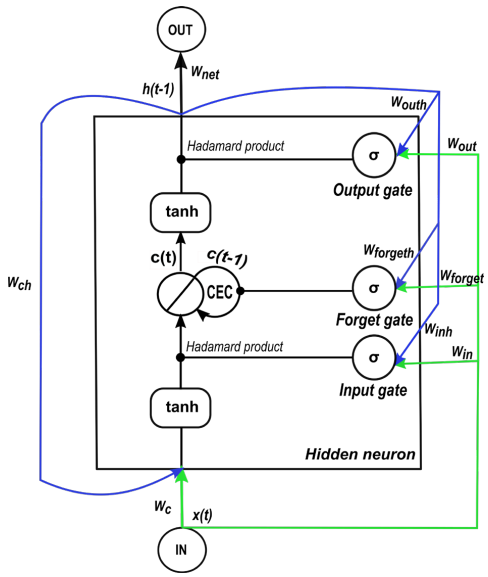

Figure 2.2: One-hidden-layer LSTM network with one hidden neuron. The green lines indicate the entry points of new inputs into the hidden neuron. The blue lines show the entry points of previous outputs into the hidden neuron, where w_* is the weight on a linkage, $h(*)$ is the output of the hidden neuron, $x(t)$ is the input at the time step t, and $c(*)$ is the cell state of the CEC. σ represents a sigmoid function of a gate, and tanh is a hyperbolic tangent function.

The procedure for processing inputs in hidden neurons of LSTM networks is as follows (Olah, 2015; Ma et al., 2019): 1) filter the information used for prediction from new inputs based on the result of the input gate, 2) filter the information to be remembered from the old CEC state according to the output of the forget gate, 3) update the CEC state using the results from the previous two steps, and 4) compute outputs of hidden neurons from the new CEC state and the results given by the output gate.

Figure 2.2 illustrates a one-hidden-layer LSTM network containing only one hidden neuron; the pseudocode is presented in Appendix A to detail how data are transferred in the given LSTM network. Owing to the limited data available at each pixel (i.e., a total of 252 time steps), we built small LSTM networks at the local

scale with one input layer, one hidden layer, and one output layer. The network receives monthly pr_a from the input layer, processes it on the hidden layer, and finally generates monthly wtd_a from the output layer. The numbers of input and output neurons are determined by how many input and output variables are used in the derivation of the network. In the constructed LSTM networks, only one neuron is located on either the input or output layer, as the number of input or output variables is one. Thus, the complexity of the network only depends on the number of hidden neurons and, therefore, can vary by changing the number of hidden neurons. The architecture of a network plays an important role in its behavior when processing new data, and it can be a double-edged sword to apply a network with considerable hidden neurons. On the one hand, the bigger we allow a network to grow, the better it can learn from a given data set. On the other hand, a complex network easily captures unwanted patterns when it learns too much from the given data set, eliminating its ability to deal with previously unobserved information (Dawson and Wilby, 2001; Müller and Guido, 2017b). This phenomenon is termed overfitting. Hence, it is crucial to identify the optimal number of hidden neurons and specify the appropriate structure of the network, which is the focus of the hyperparameter tuning described in Section 2.2.5.

2.2.3 Continuous and cross-wavelet transform

Continuous wavelet transform (CWT) is a type of wavelet transform useful for feature extraction (Grinsted et al., 2004). Given a mother wavelet $\psi_0(\eta)$, with η being a dimensionless time parameter, the CWT of a time series x_{n_0} is formulated as the convolution of x_{n_0} and a scaled and translated form of $\psi_0(\eta)$ as follows (Torrence and Compo, 1998):

$$W(s,n) = \sum_{n_0=0}^{N-1} x_{n_0} \psi^*[(n_0 - n)\delta t/s] \,, \tag{2.3}$$

where the asterisk (*) signifies the complex conjugate, δt is the time step of x_{n_0}, N is the total number of δt in x_{n_0}, s is the wavelet scale, and n is the localized time index along which $\psi_0(\eta)$ is translated. Here, the wavelet power is defined as $|W(s,n)|^2$.

The mother wavelet must be zero mean and localized in the time and frequency domains (Torrence and Compo, 1998). In this study, we applied the Morlet wavelet as the mother wavelet, defined as follows:

$$\psi_0(\eta) = \pi^{-1/4} e^{i\omega_0 \eta} e^{-\eta^2/2} \,, \tag{2.4}$$

where ω_0 is the dimensionless frequency, which is set as 6 here to acquire a good balance between time and frequency localization (Grinsted et al., 2004).

XWT is a method for locating common high power in the wavelet transforms of two time series. The XWT of two time series x_{n_0} and y_{n_0} can be computed using the following (Grinsted et al., 2004):

$$W_{xy}(s,n) = W_x(s,n) W_y^*(s,n) \,, \tag{2.5}$$

where $W_x(s,n)$ and $W_y(s,n)$ are the CWT of the time series x_{n_0} and y_{n_0}, respectively. The cross-wavelet power is calculated as $|W_{xy}(s,n)|$. However, directly using the cross-wavelet power gives biased results of the XWT analysis, so we applied $|W_{xy}(s,n)|/s$ proposed by Veleda et al. (2012) for correction. For detailed descriptions about CWT and XWT, the reader is referred to Torrence and Compo (1998), Grinsted et al. (2004), Prokoph and El Bilali (2008), and Veleda et al. (2012).

In this study, XWT is used as an independent and additional analysis tool to visualize the pattern in the pr_a-wtd_a relationship at the individual pixel level in time and frequency. In the XWT analyses, we focus on common, localized high-power frequency modes of $\psi_0(\eta)$ in pr_a and wtd_a time series and the dynamics of the modes over time. Using the XWT analysis, we expect to clarify whether a changing pattern exists in the pr_a-wtd_a relationship during the study period and if it affects the network behavior. Moreover, by linking the results of the XWT analysis with the network outputs, we explore the impact of the amount and range of the frequency modes on the LSTM network performance in order to obtain insight into the internal operations of LSTM networks.

2.2.4 Study area and data set

We constructed the LSTM networks at individual pixels over eight hydrometeorologically different regions within Europe (Figure 2.3), which are known as the PRUDENCE (Prediction of Regional scenarios and Uncertainties for Defining EuropeaN Climate change risks and Effects) regions (Christensen and Christensen, 2007). Table 2.1 lists the region names and abbreviations, coordinates, and climatologic information. The climatology is represented by regional averages and standard deviations of yearly averaged data derived from the TSMP-G2A data set (Furusho-Percot et al., 2019) from the years 1996 to 2016, except for *SWE* of which data are only available from the years 2003 to 2010. The TSMP-G2A data set consists of daily averaged simulation results from TSMP over Europe, using the grid definition from the COordinated Regional Downscaling Experiment (CORDEX) framework with a spatial resolution of 0.11° (12.5 km; EUR-11). TSMP is a fully coupled atmosphere-land-surface-subsurface modeling system, giving a physically consistent representation of the terrestrial water and energy cycle from the GW via the land surface to the top of the atmosphere, which is unique (Keune et al., 2016; Furusho-Percot et al., 2019). The current version (version 1.1) of TSMP consists of the numerical weather prediction model of COnsortium for Small-scale MOdeling (COSMO), version 5.01, the National Center for Atmospheric Research (NCAR) Community Land Model (CLM), version 3.5, and the 3D surface–subsurface hydrologic model ParFlow, version 3.2, which are externally coupled by the Ocean Atmosphere Sea Ice Soil Model Coupling Toolkit (OASIS3–MCT) coupler (Gasper et al., 2014; Shrestha et al., 2014). TSMP has been successfully applied in many studies to simulate the terrestrial hydrological processes (e.g., Shrestha et al., 2014; Kurtz et al., 2016; Sulis et al., 2018; Keune et al., 2019). Furusho-Percot et al. (2019) showed good agreement of the hydrometeorological variability between TSMP-G2A and observed data at the regional scale in the PRUDENCE regions. They compared anomalies of temperature, precipitation, and total column water storage from the TSMP-G2A data set with commonly used reference observational datasets (i.e., the 0.25

degrees gridded European Climate Assessment and Dataset, ECA&D, E-OBS v19, and observations from the Gravity Recovery and Climate Experiment, GRACE), resulting in Pearson correlation coefficients (r) ranging from 0.73 to 0.94 for temperature anomalies and from 0.62 to 0.88 for pr_a. Similar results were obtained by Hartick et al. (2021), who compared anomalies of total column water storage from the TSMP-G2A data set with the novel GRACE-REC data set and obtained r from 0.69 to 0.89 in the different PRUDENCE regions. For details of the TSMP-G2A data set, the reader is referred to Furusho-Percot et al. (2019).

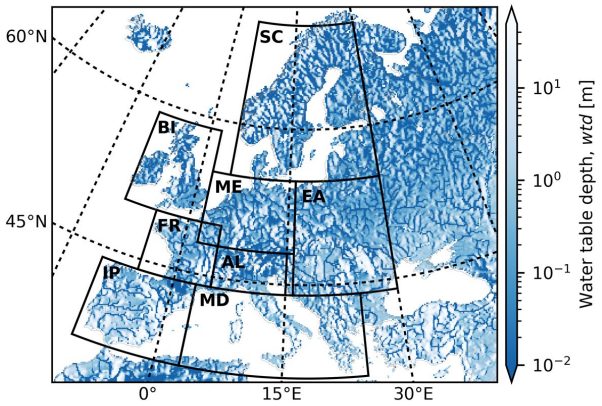

Figure 2.3: TSMP-G2A *wtd* [m] climatology over the European continent for the time period from 1996 to 2016. Areas bounded by the thick black lines show the PRUDENCE regions (i.e., SC - Scandinavia; BI - British Isles; ME - Mid-Europe; EA - Eastern Europe; FR - France; AL - Alps; IP - Iberian Peninsula; MD - Mediterranean).

Table 2.1: Overview of the PRUDENCE regions, including region names and abbreviations, coordinates, and climatologic information extracted from the TSMP-G2A data set (expressed as average ± standard deviation).

Area	Coordinate (long_west, long_east, lat_south, lat_north)	Regional precipitation [mm]	Regional water table depth, *wtd* [m]	Regional evapotranspiration, *ET* [mm]	Regional soil moisture, θ [$m^3 m^{-3}$]	Regional snow water equivalent, *SWE* [mm]
Scandinavia (SC)	(5, 30, 55, 70)	1005 ± 451	2.43 ± 5.83	283 ± 129	0.32 ± 0.11	79.80 ± 109.17
British Isles (BI)	(-10, 2, 50, 59)	1119 ± 308	2.29 ± 6.11	395 ± 130	0.36 ± 0.10	0.82 ± 2.19
Mid-Europe (ME)	(2, 16, 48, 55)	885 ± 192	2.77 ± 6.87	444 ± 141	0.35 ± 0.10	2.44 ± 5.49
Eastern Europe (EA)	(16, 30, 44, 55)	776 ± 185	3.08 ± 7.37	470 ± 164	0.33 ± 0.10	9.50 ± 13.07
France (FR)	(-5, 5, 44, 50)	897 ± 169	2.95 ± 7.04	485 ± 164	0.35 ± 0.10	0.31 ± 1.12
Alps (AL)	(5, 15, 44, 48)	1494 ± 638	4.14 ± 9.16	499 ± 185	0.35 ± 0.10	65.57 ± 127.23
Iberian Peninsula (IP)	(-10, 3, 36, 44)	841 ± 371	6.62 ± 10.61	495 ± 233	0.29 ± 0.11	3.38 ± 28.18
Mediterranean (MD)	(3, 25, 36, 44)	894 ± 338	6.48 ± 10.76	518 ± 229	0.30 ± 0.10	3.59 ± 15.22

As shown by the averages in Table 2.1, precipitation is heterogeneously distributed over the PRUDENCE regions, with the highest rainfall in AL (1494 mm) and the lowest in EA (776 mm). Most regional average

wtd's range from 2 m to 5 m, other than IP and MD (having a larger average *wtd* > 6 m). Within this range, AL has a relatively high average *wtd* (4.14 m) due to its strong relief. Higher *ET* is naturally observed in more arid regions, e.g., the highest regional average *ET* (518 mm) is recorded in MD. No significant difference is observed in regional average θ over PRUDENCE regions, and the minimal regional average θ is observed in IP (0.29 m^3m^{-3}) and MD (0.30 m^3m^{-3}). For *SWE*, large values (> 60 mm) are simulated in SC and AL, while values below 10 mm are recorded in the other regions.

We utilized the TSMP-G2A data set to compute pr_a and wtd_a (Eqs. 2.6 and 2.7) at the individual pixel level over Europe, which are the input and output data of the proposed LSTM networks. The associated average and standard deviation values are based on the training set (i.e., the data within the years 1996 to 2012, described in Section 2.2.5) to guarantee that no future information leaks into the networks in the training process.

$$pr_a = (pr_m - pr_{av})/pr_{sd} \,, \tag{2.6}$$

where pr_m is the monthly sum precipitation calculated from the TSMP-G2A data set, pr_{av} is the climatological average of pr_m (i.e., averages of pr_m in January, February, …, December), and pr_{sd} is the climatological standard deviation of pr_m.

$$wtd_a = (wtd_m - wtd_{av})/wtd_{sd} \,, \tag{2.7}$$

where wtd_m is the monthly average *wtd* derived from the TSMP-G2A data set, wtd_{av} is the climatological average of wtd_m, and wtd_{sd} is the climatological standard deviation of wtd_m.

The wtd_a is a measure of GW drought. Here we define $wtd_a \geq 2$ corresponding to extreme drought, $1.5 \leq wtd_a < 2$ corresponding to severe drought, $1 \leq wtd_a < 1.5$ corresponding to moderate drought, $0 \leq wtd_a < 1$ corresponding to minor drought and $wtd_a < 0$ corresponding to no drought, following McKee et al. (1993).

To identify the effect of local factors on the network behaviors, we categorized the network performances based on different intervals of yearly averaged *wtd*, *ET*, θ, *SWE*, and S_t and dominant *PFT*. The data of θ were calculated based on the information at a depth from 0 to 5 cm below the land surface. It is important to note that the data used in this study cover the years 1996 to 2016 (except for *SWE* data, which are only available from 2003 to 2010) to ensure that spinup effects do not impact the analyses (Furusho-Percot et al., 2019).

2.2.5 Experiment design

LSTM networks are employed here to detect connections between pr_a and wtd_a from the pan-European simulation results and utilize pr_a as input to predict wtd_a. At each time step, one new input enters a network, together with information stored in the network's memory (i.e., useful messages from inputs in the past) to generate outputs. Therefore, LSTM networks have the ability to handle the lagged response of wtd_a to pr_a.

Monthly anomaly time series at individual pixels were divided into three parts for network training (01/1996–12/2012), validation (01/2013–12/2014), and testing (01/2015–12/2016) containing about 80%, 10%, and 10% of the total data, respectively. In training, the network is fitted to a given training set by adjusting its weights and biases. The technique of adjusting network parameters is called an optimizer that minimizes a cost function at a certain learning rate (Govindaraju, 2000). This study utilized a supervised training algorithm with a supplementary teacher signal (i.e., TSMP-G2A monthly wtd_a) to guide the training process, which is widely adopted in hydroscience in case of, e.g., stream stage modeling (Sung et al., 2017), stream discharge modeling (Zhang et al., 2015), and GW level modeling (Adamowski and Chan, 2011). One common challenge in the training process is overfitting. Validation is a process to address overfitting by comparing the network output with the teacher signal to obtain a validation loss (Govindaraju, 2000; Liong et al., 2000). Provided that the network has gained sufficient knowledge from the training set, training ceases when the number of epochs (i.e., an iteration when the whole training set travels through the network forward and backward once) is ≥ 50, and the validation loss starts increasing. The strategy to stop training based on validation losses is termed early stopping.

Moreover, the validation losses were applied to tune hyperparameters of the LSTM networks whose architecture was introduced in Section 2.2.2. To simplify the procedure of hyperparameter tuning, we only focused on the optimization of the number of hidden neurons in this study and set other hyperparameters to be constant (Table 2.2). The networks with hidden neurons from 1 to 100 were trained at individual pixels, and the best three of them were selected for testing based on the validation losses.

Table 2.2: Hyperparameter settings of the proposed LSTM networks.

Hyperparameter	Value or method
Number of input, hidden, and output layer(s)	(1, 1, 1)
Number of input, hidden and output neuron(s)	(1, 1-100, 1)
Initial weights and biases of all neurons	U(-0.5, 0.5)*
Initial cell states of LSTM neurons	0
Optimizer and learning rate	RMSprop (Hinton et al., 2012), 0.001
Loss function	Mean square error (MSE)

* U(-0.5, 0.5) - uniform distribution bounded by -0.5 and 0.5.

Finally, during testing, the optimally trained networks were provided with a previously unknown data set, originating from the same source as the training set. The difference between generated and target values during testing is called the generalization error, representing the ability of a network to perform on previously unobserved data. The average of the three optimal network results was utilized for evaluation in order to moderately eliminate the individual deficiencies of the selected networks, thereby improving the quality of the final results (Goodfellow et al., 2017b; Brownlee, 2018). The metrics to assess network performance in this study are the root mean square error (RMSE), the Nash-Sutcliffe efficiency (NSE) and the bias from r (α) as shown in Eqs. (2.8)-(2.10), respectively. α indicates systematic additive and multiplicative biases in the generated values, with a value between 0 and 1, where α = 1 means no bias (Duveiller et al., 2016).

$$RMSE = \sqrt{\sum_{i=1}^{N}(y_t - y_e)^2 / N}\,, \tag{2.8}$$

$$NSE = 1 - \sum_{i=1}^{N}(y_t - y_e)^2 / \sum_{i=1}^{N}(y_t - \overline{y_t})^2, \tag{2.9}$$

$$\alpha = 2 / \left[\sigma_{y_t}/\sigma_{y_e} + \sigma_{y_e}/\sigma_{y_t} + (\overline{y_t} - \overline{y_e})^2/(\sigma_{y_t}\sigma_{y_e})\right], \tag{2.10}$$

where y_t, $\overline{y_t}$, and σ_{y_t} are the target value, the average of the target values, and the standard deviation of the target values, respectively. y_e, $\overline{y_e}$, and σ_{y_e} are the estimated value, the average of the estimated values, and the standard deviation of the estimated values, respectively. N is the number of time steps in the given time series.

Repeating the above network training, validation, and testing processes (right panel of Figure 2.4), we constructed the proposed LSTM networks locally at ≤ 200 pixels randomly selected in each group in order to save computing time. As described in Section 2.2.4, climatologic differences occur not only between different PRUDENCE regions but also at certain pixels in the same region, which potentially explains varying network performances at individual pixels. To analyze the network reaction to local factors, we categorized the pixels into groups based on various intervals of yearly averaged *wtd*, *ET*, *θ, SWE*, and S_t and dominant *PFT* (Table 2.3), and the analysis results will be presented in Section 2.3.2. Figure 2.4 gives a generic workflow of this study to establish the LSTM networks at the local scale and analyze their output.

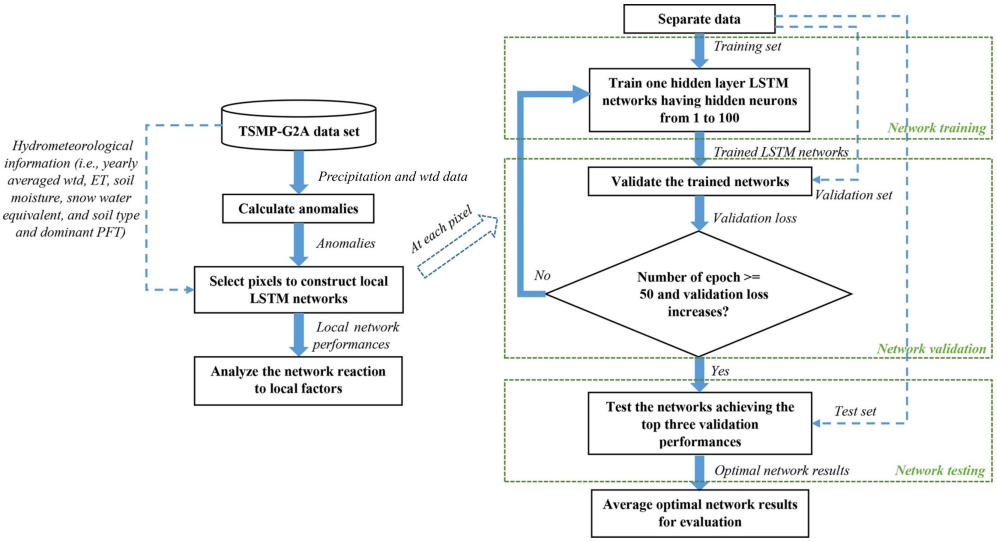

Figure 2.4: Workflow for LSTM network setup over the European CORDEX domain. The left section represents the overall processes of the network setup, whereas the right section shows how to apply LSTM networks at a selected pixel. The blue dashed lines with arrows indicate additional data transmission paths.

Table 2.3: Intervals of yearly averaged *wtd*, *ET*, θ, *SWE*, and S_t and dominant *PFT* for categorization.

Yearly averaged water table depth, *wtd* [m]	Yearly averaged evapotranspiration, *ET* [mm]	Yearly averaged soil moisture, θ [$m^3\ m^{-3}$]	Yearly averaged snow water equivalent, *SWE* [mm]	Soil type, S_t	Dominant plant functional type, *PFT**
0.0–1.0	< 0.0	0.0-0.05	≤ 10.0	Sand	Needleleaf evergreen temperate tree
1.0-2.0	0.0-100.0	0.05-0.10	> 10.0	Loamy sand	Needleleaf evergreen boreal tree
2.0-3.0	100.0-200.0	0.10-0.15		Sandy loam	Needleleaf deciduous boreal tree
3.0-4.0	200.0-300.0	0.15-0.20		Silt loam	Broadleaf evergreen tropical tree
4.0-5.0	300.0-400.0	0.20-0.25		Silt	Broadleaf evergreen temperate tree
5.0-6.0	400.0-500.0	0.25-0.30		Loam	Broadleaf deciduous tropical tree
6.0-7.0	500.0-600.0	0.30-0.35		Sandy clay loam	Broadleaf deciduous temperate tree
7.0-8.0	600.0-700.0	0.35-0.40		Silty clay loam	Broadleaf deciduous boreal tree
8.0-9.0	700.0-800.0	0.40-0.45		Clay loam	Broadleaf evergreen shrub
9.0-10.0	800.0-900.0	0.45-0.50		Sandy clay	Broadleaf deciduous temperate shrub
10.0-50.0	900.0-1000.0			Silty clay	Broadleaf deciduous boreal shrub
	1000.0-1100.0			Clay	C3 Arctic grass
				Organic Material	C3 non-Arctic grass
				Water	C4 grass
				Bedrock	Corn
				Others	Wheat

*Dominant *PFT* - the *PFT* of which percentage is ≥ 50% at one pixel.

2.3 Results and discussion

2.3.1 Water table depth anomaly maps in 2003 and 2015 reproduced by the LSTM network results

We employed the outputs of the proposed LSTM networks to reproduce wtd_a over the European continent in 2003 and 2015, constituting drought years (Van Loon et al., 2017). Here, we displayed the wtd_a from the TSMP-G2A data set and the networks (hereafter called LSTM wtd_a) for August 2003 (in the training period; Figure 2.5a) and August 2015 (in the testing period; Figure 2.5b) with respect to strength. We focused on areas where $wtd_a \geq 1.5$ (i.e., a strong drought) and studied the consistency between the TSMP-G2A and the LSTM wtd_a results on the distribution of GW drought. As the LSTM networks performed well at most pixels during the training period, the LSTM wtd_a map appears almost identical to the TSMP-G2A wtd_a map for August 2003 (see Figure 2.5a), showing severe GW drought in most parts of Europe, which is in good agreement with previous studies (Andersen et al., 2005; Van Loon et al., 2017). Moreover, in the simulations and LSTM results, there is increased GW storage over central Germany, central Britain, southeastern France, the western Iberian Peninsula, and several parts in Eastern Europe, illustrating the strong spatial heterogeneity of the anomalies, which is expected. In contrast, due to decreased network performance during testing, the LSTM wtd_a map shows less agreement with the TSMP-G2A wtd_a map for August 2015 (see Figure 2.5b) with respect to the severity of drought. Extremes in wet and dry anomalies (i.e., $|wtd_a| \geq 2$) were especially underestimated, suggesting that the training set contains too little information on extreme

events and, thus, is too short. Yet overall, a visual inspection of Figure 2.5b shows that the LSTM wtd_a map agrees well with the TSMP-G2A wtd_a map on the spatial distribution of dry and wet events. In both maps, we identified severe drought in mid-Europe, the Alps, and the northwest of Eastern Europe, lending confidence in the trained networks to predict wtd_a from pr_a information. Additional European wtd_a maps for the second half of 2003 and 2015 are shown in Figures A.1 and A.2, leading to similar conclusions regarding the ability of the LSTM results to reproduce TSMP-G2A wtd_a.

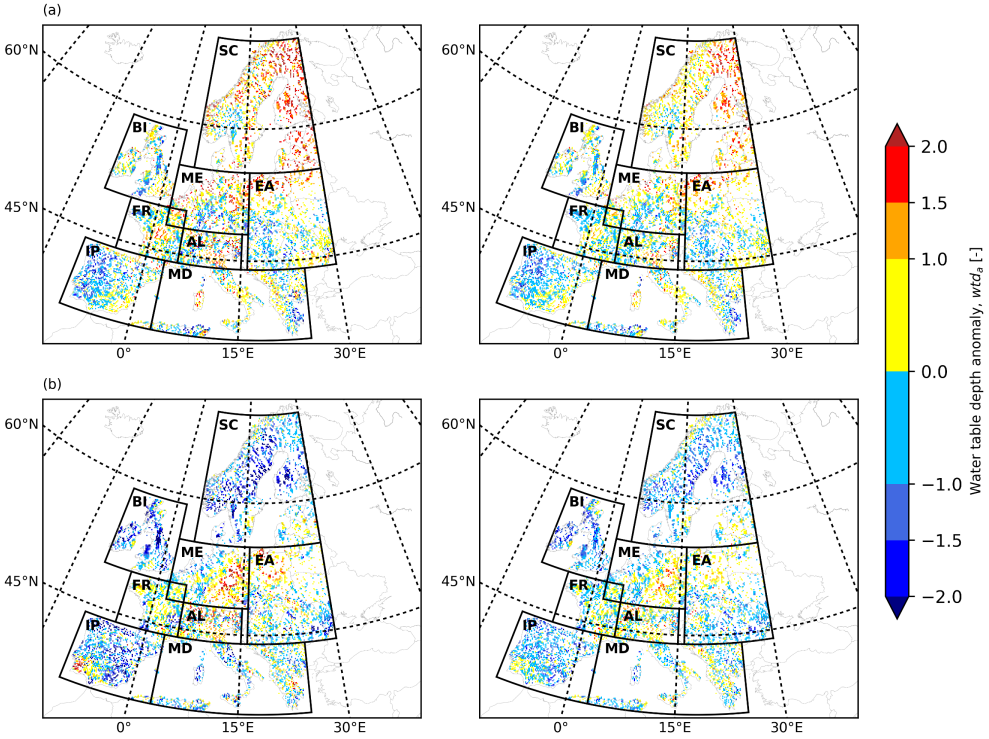

Figure 2.5: European wtd_a maps for (a) August 2003 (i.e., in the training period) and (b) August 2015 (i.e., in the testing period), derived from the TSMP-G2A data set (left) and the results from LSTM networks (right).

2.3.2 Impact of local factors on the network performance

In each PRUDENCE region, we computed the averages and standard deviations of the test NSE scores and RMSEs for the categories based on different intervals (Table 2.3) of yearly averaged *wtd*, *ET*, *θ, SWE*, and S_t and dominant *PFT* (Figure 2.6) to study dependents of the network test performances on different local factors. For statistical significance, we only considered categories with ≥ 50 pixels. In addition, negative NSE values at the pixel level were set to zero in the calculation of averages and standard deviations.

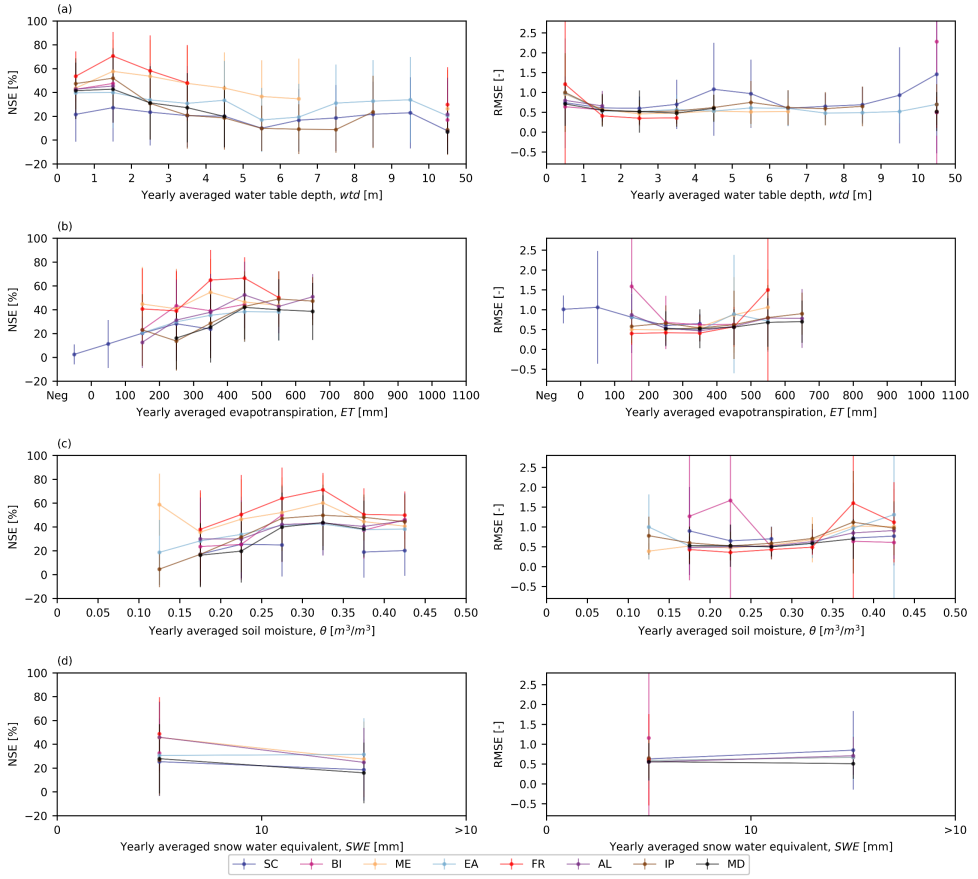

Figure 2.6: Averages and standard deviations of the test NSE scores (left) and RMSEs (right) over the categorized results. Shown are the yearly averaged (a) *wtd*, (b) *ET*, (c) *θ*, and (d) *SWE*. The averages are indicated as dots, while the bars indicate standard deviations. The different colors reflect test results in different PRUDENCE regions.

There was no significant influence of S_t and dominant *PFT* on the scores (not shown here). In general, the performance decreased with increasing yearly averaged *wtd*, which was manifested by decreasing average NSE scores and growing average RMSEs (Figure 2.6a). This type of network behavior can be attributed to a stronger connection of GW to precipitation in shallow aquifers, which is intuitive. In contrast to the impact of yearly averaged *wtd* on the test performance, the performance was positively correlated to yearly averaged *ET* and *θ*. With increasing yearly averaged *ET* (Figure 2.6b) or *θ* (Figure 2.6c), there was an increase in average NSE scores and a decrease in average RMSEs. We can explain this phenomenon by the overlap between low-*wtd* and high-*ET* (or high-*θ*) areas over Europe. We also discovered that yearly averaged *SWE* played an important role in the network test performance. In most PRUDENCE regions, the performance decreased in the case of increasing *SWE*, leading to smaller average NSE scores and larger

average RMSEs presented in Figure 2.6d. The reason is that snow accumulation resulted in complex feedback with GW processes that cannot be captured well by the networks without including additional input information. Overall, in the same region, most of the proposed LSTM networks achieved a relatively good test performance at the pixels with yearly averaged wtd < 3 m, ET > 200 mm, θ > 0.15 m^3m^{-3} and SWE < 10 mm, where a stronger relationship exists between pr_a and wtd_a.

As mentioned in Section 2.2.4, we only used the training set to calculate the climatological average and standard deviation in order to prevent the networks from incorporating future information in the training process. However, some extreme values in the validation and test sets may exceed the range of the training set, resulting in decreased validation and test performances and suggesting that a varying pattern may exist between pr_a and wtd_a over the study period (see discussion in Section 2.3.3). This can also be a potential reason for large standard deviations in the test RMSEs in Figure 2.6.

Figure 2.6 also reveals different regional network test performances. In the same interval of yearly averaged wtd, the difference in yearly average NSE scores between two PRUDENCE regions can be more than 40%. FR exhibits the overall best network performance during testing. As shown in Table 2.1, the regional average wtd, ET, θ, and SWE of FR are 2.95 m (< 3 m), 485 mm (> 200 mm), 0.35 m^3m^{-3} (> 0.15 m^3m^{-3}), and 0.31 mm (< 10 mm), respectively. Hence, there was a close connection between pr_a and wtd_a at most pixels in FR, resulting in good network test performance.

To further analyze the network test performances in different PRUDENCE regions, Figure 2.7 and Table 2.4 provide test NSE scores over Europe and percentages of the selected pixels with test NSE ≥ 50%, respectively. FR outperformed the other regions on test NSE scores, which is consistent with the finding from Figure 2.6. In BI, ME, EA and AL, the proposed LSTM networks behaved well during testing (i.e., having test NSE ≥ 50%) at more than 30% of the selected pixels (colored in blue in Figure 2.7). However, we also found low percentages of the selected pixels with test NSE ≥ 50% in SC, IP, and MD, which are 17.46%, 29.66%, and 27.72%, respectively. In Table 2.1, SC is characterized as the region with the largest regional average SWE (79.80 mm) and the smallest regional average ET (283 mm), and as shown in Figure 2.6, the networks tended to perform poorly during testing in the areas with large SWE and small ET. The pixels in IP and MD (regional average wtd > 6 m) generally have larger wtd than the other regions, resulting in a more lagged and weaker connection between pr_a and wtd_a, which is intuitive. Therefore, the network behavior in IP and MD was relatively poor.

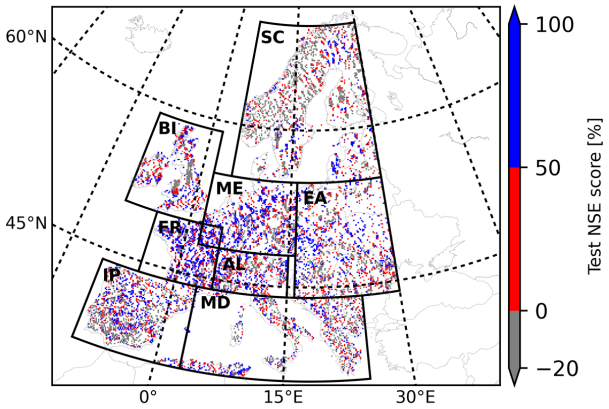

Figure 2.7: Map of test NSE scores achieved by the proposed LSTM networks in the PRUDENCE regions.

Table 2.4: Percentages of the selected pixels with a test NSE score ≥ 50% in the PRUDENCE regions [%].

SC	BI	ME	EA	FR	AL	IP	MD
17.46	33.38	50.00	32.09	57.28	40.06	29.66	27.72

We extended the scope of the analyses to the entire study period and found that the performance of individual networks generally followed two combinations with respect to training and test scores, which are as follows:

- C1: training NSE score ≥ 50%, test NSE score ≥50%;
- C2: training NSE score ≥ 50%, test NSE score ≤ 0%.

The data distribution in the training and test sets was expected to be analogous, and if the networks did not encounter overfitting during training, their test performance increased by the improvement of the training performance, and vice versa. C1 is the expected network behavior with both satisfactory training and test scores. C2 is an exception in which the networks that performed well on the training set failed to handle the test set. Significantly reduced test performance in C2 can be attributed to the hypothesis that the pattern between pr_a and wtd_a varied over the study period.

Figure 2.8 shows the percentages of the pixels where the network performance followed C1 (Figure 2.8a) and C2 (Figure 2.8b) in different PRUDENCE regions and intervals of *wtd*, *ET*, *θ and SWE*. For statistical significance, only the regions and the intervals of *wtd*, *ET*, *θ and SWE* with ≥ 50 selected pixels were considered. Here, we focused on the regions and the intervals with high percentages (> 30%; above the black dashed lines in Figure 2.8) to identify common hydrometeorological characteristics of a pixel where the network performance followed C1 or C2. For C1, high percentages were found in regions except for SC, IP and MD and in areas with *wtd* ≤ 3 m, *ET* ≥ 200 mm, $\theta \leq 0.10$ m^3m^{-3} and $\theta \geq 0.20$ m^3m^{-3}, and *SWE* ≤ 10 mm, which are in good agreement with our previous findings. In contrast, for C2, the percentages are high

in SC, EA, IP and MD and in areas with $wtd \geq 2$ m, $ET \leq 300$ mm, and 0.10 $m^3m^{-3} < \theta \leq 0.25$ m^3m^{-3}. The distribution of C2 is not very sensitive to *SWE*, and the percentages are large in both areas with $SWE \leq 10$ mm and $SWE > 10$ mm. Moreover, in areas with negative *ET*, there is no pixel where the network performances followed C1, and C2 is the dominant network performance combination. We explain negative *ET* by pronounced freezing and sublimation processes in these areas, which significantly affect the response of wtd_a to pr_a.

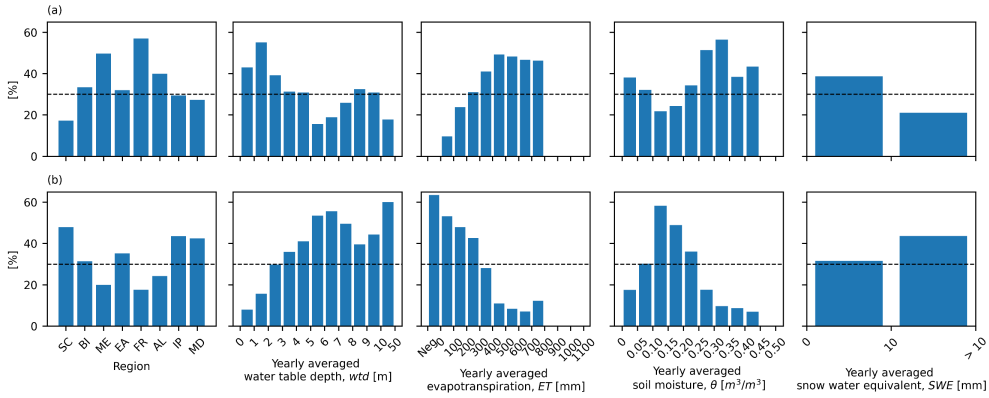

Figure 2.8: Bar plots showing the percentages of pixels where the network performance followed the combinations (a) C1, and (b) C2 in different regions and intervals of yearly averaged *wtd*, *ET*, θ, and *SWE*, from left to right, respectively. Black dashed lines indicate percentages equal to 30%.

2.3.3 Cross-wavelet transform (XWT) analysis

In the previous section, we posed the hypothesis that the temporal pattern between pr_a and wtd_a during training, validation, and testing was different at a number of pixels over the European continent. XWT was employed here for hypothesis testing at the individual, representative pixels (Table 2.5), which were randomly selected based on the hydrometeorological characteristics of C1 and C2 summarized in Figure 2.8. XWT showed the time-frequency pattern in the pr_a and wtd_a time series derived from the TSMP-G2A data set (i.e., TSMP-G2A pr_a and wtd_a) at these pixels and highlighted the common high power of the frequency components in the time series (Figure 2.9). The α values (Eq. 2.10) of pixel 1 were generally suggesting that smaller biases existed in the results of the LSTM networks. In addition, we found different α values for pixel 2 with small biases in the training and large biases in the validation and testing.

Figure 2.9 shows the results of the XWT analyses of the selected pixels in combination with the corresponding TSMP-G2A pr_a and wtd_a time series. Inspecting the results of the XWT analyses (bottom panel of Figure 2.9), the concentration period of power was inconsistent in the area without edge effects (i.e., the area within the black dashed line) at pixel 2 from the time period 1996 to 2016, indicating a time-varying pattern between pr_a and wtd_a at the pixel, thus supporting our hypothesis. It also explores the high sensitivity of LSTM networks to outliers, which is a drawback of data-driven models.

The high power in the XWT results at the representative pixel of C1 (pixel 1; Figure 2.9a) was consistently located in a certain period (i.e., below 64 months), indicating a consistent pattern between pr_a and wtd_a throughout the whole study period, which is the prerequisite of good network performances. At the pixel, we found that most of the high power in the XWT results was consistently concentrated in the period from 2 to 16 months during the study period (see Figure 2.9a). Figure A.3a showed similar phenomena as above. Therefore, we speculate that LSTM networks might be frequency aware and work well to capture the pr_a-wtd_a relationship at the monthly, seasonal and annual periods.

Table 2.5: Pixel characteristics in the XWT analysis (pixels 1-2).

	Performance combination	Region	Yearly averaged water table depth, *wtd* [m]	Yearly averaged evapotranspiration, *ET* [mm]	Yearly averaged soil moisture, θ [m^3 m^{-3}]	Yearly average snow water equivalent, *SWE* [mm]
Pixel 1	C1	FR	1.38	422.91	0.28	0.0
Pixel 2	C2	SC	5.19	-24.41	0.16	535.0
	Training NSE [%]	Training α [%]	Validation NSE [%]	Validation α [%]	Test NSE [%]	Test α [%]
Pixel 1	82.50	98.94	53.99	91.32	82.63	99.09
Pixel 2	66.47	93.72	-34.74	34.82	-802.83	8.84

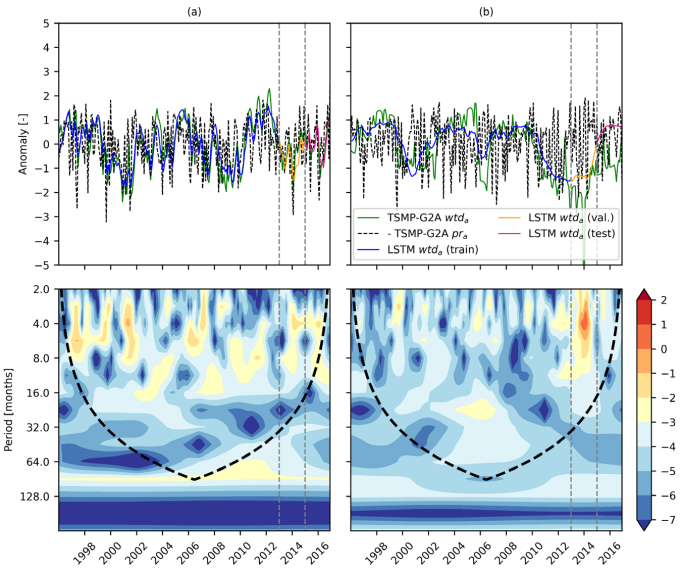

Figure 2.9: TSMP-G2A pr_a, TSMP-G2A wtd_a, and LSTM wtd_a time series (top), as well as cross-wavelet spectra for TSMP-G2A pr_a and wtd_a series (bottom), at a representative pixel of the performance combination (a) C1 and (b) C2. In the cross-wavelet spectra, the black dashed line marks the boundary of the cone of influence. The color bar presents $\log_2(\text{power/scale})$. In all plots, the two gray dashed lines separate the study period into the training, validation, and testing periods.

2.4 Summary and conclusions

In this study, we proposed LSTM networks as an indirect method to model monthly wtd_a over the European continent, using monthly pr_a as input. Local LSTM networks were constructed at individual pixels randomly selected over Europe to capture the time-varying and time-lagged relationship between pr_a and wtd_a from integrated hydrologic simulation (TSMP-G2A) results covering the 1996 to 2016 episode. The monthly anomaly series derived from the TSMP-G2A data set were divided into three sections at each pixel for network training, validation, and testing. Using the output of the LSTM networks, we successfully reproduced TSMP-G2A wtd_a maps over Europe for drought months in both the training and testing period (e.g., August 2003 and August 2015) in terms of the spatial distribution of dry and wet events. The good agreement between the TSMP-G2A and LSTM wtd_a maps demonstrated the ability of the trained networks to model wtd_a from pr_a data. The results highlighted the impact of local factors on the network test performance, as manifested by NSE scores and RMSEs. Most of the networks attained high test NSE scores at the pixels with wtd < 3 m, ET > 200 mm, θ > 0.15 m^3m^{-3}, and SWE < 10 mm, where a stronger connection existed between pr_a and wtd_a. Also, the various hydrometeorological characteristics in each PRUDENCE region resulted in regional differences in the test performance of the proposed networks, with FR showing the overall best network performance. In some regions, test performance deteriorated due to changing temporal patterns in the pr_a-wtd_a relationship, approved by XWT analyses. According to the results of the XWT analyses, we hypothesize that LSTM networks have frequency awareness and tend to perform well to capture the pr_a-wtd_a relationship at the monthly, seasonal and annual periods.

We also recognized that the limited amount of data in the training introduces uncertainties in the network performances. Any potential extension of training data may lead to a significant improvement in the quality of the derived networks. In addition, hyperparameters of the proposed LSTM networks may be further tuned at the individual pixel level to improve network performance. Due to a lack of spatiotemporally continuous wtd observations over Europe, this study presents a methodology for deriving a LSTM network model for wtd_a from pr_a based on simulation results from a terrestrial model (i.e., the TSMP-G2A data set). As demonstrated in Furusho-Percot et al. (2019) and Hartick et al. (2021), the TSMP-G2A data set shows a good agreement with hydrometeorological and GRACE observations in different European regions. Therefore, we argue that the TSMP-G2A data set is a good reference data set to establish the methodology. The results suggest that LSTM networks are useful for estimating wtd_a time series based on other hydrometeorological variables which are routinely measured and, therefore, are more easily available from, e.g., atmospheric reanalyses and forecast data sets and observations than GW level measurements. After training, LSTM networks could provide fast and reliable predictions of wtd_a only based on the data of input variables, which is impossible for traditional physically-based models such as TSMP. The proposed methodology may be transferred into a real-time monitoring and forecasting workflow for wtd_a at the continental scale.

Chapter 3 An indirect approach based on Long Short-Term Memory networks to estimate groundwater table depth anomalies across Europe with an application for drought analysis

The content of this chapter is adapted from:

Ma, Y., Montzka, C., Bayat, B. and Kollet, S.: An Indirect Approach Based on Long Short-Term Memory Networks to Estimate Groundwater Table Depth Anomalies Across Europe With an Application for Drought Analysis, Front. Water, 3, doi:10.3389/frwa.2021.723548, 2021.

Abstract

The lack of high-quality continental-scale water table depth (*wtd*) observations necessitates developing an indirect method to produce reliable estimation for water table depth anomalies (wtd_a) over Europe to facilitate European groundwater (GW) management under drought conditions. Long Short-Term Memory (LSTM) networks are a deep learning technology to exploit long-short-term dependencies in the input-output relationship, which have been observed in the response of GW dynamics to atmospheric and land surface processes. Here, we introduced different input variables including precipitation anomalies (pr_a), which is the most common proxy of wtd_a, for the networks to arrive at improved wtd_a estimates at individual pixels over Europe in various experiments. All input and target data involved in this study were obtained from the simulated TSMP-G2A data set. We performed wavelet coherence analysis to gain a comprehensive understanding of the contributions of different input variable combinations to wtd_a estimates. Based on the different experiments, we derived an indirect method utilizing LSTM networks with pr_a and soil moisture anomaly (θ_a) as input, which achieved the optimal network performance. The regional medians of test NSE scores and RMSEs obtained by the method in the areas with $wtd \leq 3.0$ m were 76%-95% and 0.17-0.30, respectively, constituting a 20%-66% increase in median NSE and a 0.19-0.30 decrease in median RMSEs compared to the LSTM networks only with pr_a as input. Our results show that introducing θ_a significantly improved the performance of the trained networks to predict wtd_a, indicating the substantial contribution of θ_a to explain GW anomalies. Also, the European wtd_a map reproduced by the method had good agreement with that derived from the TSMP-G2A data set with respect to drought severity, successfully detecting ~41% of strong drought events ($wtd_a \geq 1.5$) and ~29% of extreme drought events ($wtd_a \geq 2$) in August 2015. The study emphasizes the importance to combine soil moisture (θ) information with precipitation information in quantifying or predicting GW anomalies. In the future, the indirect method derived in this study can be transferred to real-time monitoring of GW drought at the continental scale using remotely sensed θ and precipitation observations or respective information from weather prediction models.

3.1 Introduction

Drought is one of the major natural disasters worldwide, considerably affecting environmental, human, and economic well-being. According to a report of the European Environment Agency (EEA) in 2020, in most parts of Europe, the frequency and severity of droughts have increased since 1950 and will further increase in the future (EEA, 2020). In this context, many studies on drought have been carried out over Europe, e.g., Stagge et al. (2017), Bachmair et al. (2018), and Hänsel et al. (2019).

Mishra and Singh (2010) categorized drought into five types, namely, meteorological, hydrological, agricultural, groundwater (GW) and socio-economic drought. Except for the last type, which reflects socio-economic situations, the severity of the others can be quantified by the following standardized hydrometeorological variables: 1) precipitation anomaly (pr_a) and evapotranspiration anomaly (ET_a) for meteorological drought, e.g., the Standardized Precipitation Index (SPI) (McKee et al., 1993) and the Standardized Precipitation Evapotranspiration Index (SPEI) (Vicente-Serrano et al., 2010); 2) river stage anomaly (rs_a) and river discharge anomaly for hydrological drought, e.g., the Standardized Runoff Index (Shukla and Wood, 2008) and the Streamflow Drought Index (Nalbantis and Tsakiris, 2009); 3) soil moisture anomaly (θ_a) for agricultural drought, e.g., the Crop Moisture Index (Palmer, 1968); 4) water table depth anomaly (wtd_a) for GW drought, e.g., the Standardized Groundwater level Index (Bloomfield and Marchant, 2013) and the GRACE Groundwater Drought Index (Thomas et al., 2017). These examples are not exhaustive, providing some of the related drought indices that have been widely used for extreme event analyses.

With the advances of *in-situ* and remotely sensed observation technologies, many variables mentioned above can be monitored routinely and are also available at large scales from, e.g., atmospheric reanalysis and forecast data sets, thereby significantly supporting drought investigations. However, to date, it is still challenging to obtain high-quality spatiotemporally continuous water table depth (*wtd*) measurements over Europe for the calculation of wtd_a (Van Loon et al., 2017; Brauns et al., 2020). Thus, it is necessary to develop an indirect approach to produce reliable wtd_a estimates over Europe in order to mitigate the potential negative impact of scarce *wtd* measurements on GW management at the European scale.

Indirect methods exploit the close connection between GW drought and other types of natural drought to predict wtd_a based on additional drought-related hydrometeorological variables that have data available at the continental scale. Depending on atmospheric and land surface processes, the contributions of these variables to wtd_a are non-linearly weighted and temporally lagged, which cannot be well represented by simple techniques such as using SPI or SPEI over extended time scales (commonly 6 to 12 months) to represent wtd_a.

Long Short-Term Memory (LSTM) networks are a type of recurrent neural networks (RNNs) used in the field of deep learning. Without either subjective human annotation needed in the application of simple machine learning techniques (e.g., a predefined time lag in the response of output to input variables) or extensive

physical background knowledge required by physically-based models, they can automatically detect long-short-term dependencies between input and output sequences (Reichstein et al., 2019), which are prevalent in hydrological responses. Benefiting from this characteristic, LSTM networks have recently drawn increasing attention from researchers in the hydrological sciences, e.g., rainfall-runoff modeling, Kratzert et al. (2019); flooding forecasting, Le et al. (2019); river stage modeling, Ma et al. (2019); and GW level modeling, J. Zhang et al. (2018).

In Chapter 2, we have demonstrated the utility of LSTM networks constructed at the individual pixel level to capture the time-varying and time-lagged relationship between monthly pr_a and wtd_a derived from the TSMP-G2A data set over Europe. The pr_a is the most common proxy of wtd_a. The dataset was published by Furusho-Percot et al. (2019), consisting of daily integrated hydrologic simulation results from the Terrestrial Systems Modeling Platform (TSMP). Furusho-Percot et al. (2019) and Hartick et al. (2021) corroborated the realism of the dataset in a comparison of simulated temperature, precipitation, and total column water storage anomalies with common observational datasets (i.e., E-OBS, ERA-Interim, GRACE and GRACE-REC). With the results of the proposed LSTM networks, we successfully reproduced European TSMP-G2A wtd_a maps for drought months in 2003 and 2015, showing good agreement concerning the spatial distribution of dry and wet events. Nevertheless, we also noticed relatively poor performance of the proposed networks at some pixels, which suggested the need for additional input to improve wtd_a estimates.

Introducing additional input variables supplements the information used to estimate certain frequency components of wtd_a. However, the improvement in each frequency component is not identifiable by general evaluation metrics. In this case, wavelet coherence analysis is a useful tool. The method reveals time-frequency localized coherence between time series and thus enables the detection of transient cross-correlation for a specific frequency (Labat, 2005). Several studies, such as Lane (2007), Salerno and Tartari (2009) and Fang et al. (2015), have applied wavelet coherence analysis to gain insight into the cross-correlation between modeled and target time series in the time-frequency domain.

The objective of this study was to optimize the LSTM networks proposed in Chapter 2 to arrive at improved wtd_a estimates at the individual pixel level over Europe. In addition to pr_a, which was the original input variable, we introduced ET_a, θ_a, scaled yearly averaged snow water equivalent (SWE_{scaled}), and anomalies at adjacent pixels (e.g., rs_a) as optional input variables for the networks in various experiments. Using data from the TSMP-G2A data set, we derived an indirect method based on the LSTM networks with the optimal input variable combination to estimate wtd_a at the European scale in order to facilitate European GW management under drought conditions. General evaluation metrics (i.e., the root mean square error, RMSE and the Nash-Sutcliffe efficiency, NSE) and wavelet coherence analysis provide a comprehensive understanding of the contributions of different input variables to the explanation of GW anomalies. As such, we presented and evaluated a LSTM-based method of simulated wtd_a, which can be transferred to other data sources for observation-based estimation. To our best knowledge, this study is among the first efforts

applying LSTM networks on estimating GW dynamics at the continental scale based on information in addition to meteorological data.

3.2 Methodology

We designed experiments that introduced different input variable combinations into the LSTM networks proposed in Chapter 2, which utilized a supervised training algorithm with target data (i.e., wtd_a data from the TSMP-G2A data set, hereafter named TSMP-G2A wtd_a) to guide the training process, and conducted wavelet coherence analysis to investigate the impact of the input variable combinations on the estimation of wtd_a over Europe in the time-frequency domain. In this section, we start with a brief overview of LSTM networks and wavelet coherence analysis, then describe the study area and data, and continue with the design of the performed experiments and a generic workflow to construct the LSTM networks at individual pixels.

3.2.1 Long Short-Term Memory networks

LSTM networks were introduced by Hochreiter and Schmidhuber (1997) to solve the exploding and vanishing gradient issues in standard RNNs. A LSTM network contains one input layer for receiving inputs, one or more hidden layers for internal computation and one output layer for producing final outputs. Through loops on their hidden layer(s), previous output of each hidden neuron (i.e., information-processing units on the hidden layer) in LSTM networks flows back to all neurons on the same layer and is then combined with new input to produce new neuron output. Therefore, LSTM networks are deep in time and considered to have memory (Shen, 2018). Here, we adopted the same architecture of hidden neurons as Gers et al. (2000), constructed by three gates with different functions and a constant error carousel (CEC), as illustrated in Figure 3.1. Benefiting from the interaction of these components, LSTM networks are able to exploit dependencies over 1000 time steps, surpassing standard RNNs that only remember up to 10 previous inputs (Hochreiter and Schmidhuber, 1997; Kratzert et al., 2018). The highly improved learning ability of LSTM networks facilitates estimating wtd_a in deep aquifers where a large time lag exists in the response of GW to other drought-related hydrometeorological variables. Compared to physically-based models, LSTM networks commonly necessitate less computational time and physical background knowledge (e.g., topography). Moreover, when the LSTM networks are available, they only require the data of input variables to estimate wtd_a, which can be easily accessed from observations and reanalysis products. For details about the functions of the components in LSTM hidden neurons, the reader is referred to Hochreiter and Schmidhuber (1997) and Gers et al. (2000).

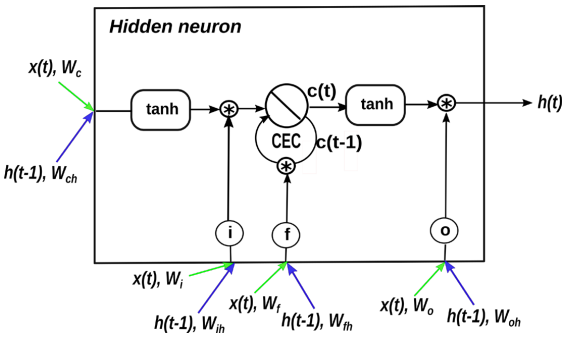

Figure 3.1: A LSTM hidden neuron. The i, f, and o represent the input, forget, and output gates, which are activated by the sigmoid function. The green arrows indicate the entry of new inputs into the hidden neuron, and the blue arrows show the entry of neuron outputs of the previous time step (i.e., t–1) into the hidden neuron. For the sake of simplicity, biases are not shown here.

Given an input variable whose time series is $x(t)$ (t ≥ 1), the computing process in a LSTM hidden neuron (Figure 3.1) at the time step t is presented by Eqs. (3.1)-(3.6).

$$i(t) = \sigma(w_i x(t) + b_i + w_{ih} h(t-1) + b_{ih}) \tag{3.1}$$

$$f(t) = \sigma(w_f x(t) + b_f + w_{fh} h(t-1) + b_{fh}) \tag{3.2}$$

$$o(t) = \sigma(w_o x(t) + b_o + w_{oh} h(t-1) + b_{oh}) \tag{3.3}$$

$$g(t) = tanh(w_c x(t) + b_c + w_{ch} h(t-1) + b_{ch}) \tag{3.4}$$

$$c(t) = f(t) * c(t-1) + i(t) * g(t) \tag{3.5}$$

$$h(t) = o(t) * \tanh(c(t)) \tag{3.6}$$

where $i(t)$, $f(t)$, $o(t)$ are the information that enters into the neuron via the input, forget and output gates, respectively. $h(t-1)$ and $h(t)$ are the neuron output at time step t - 1 and t, respectively. $c(t-1)$ and $c(t)$ are the cell state at time step t - 1 and t, respectively. w_* and b_* are the learnable weight and bias on a linkage between neurons, respectively. The subscripts i, f, o and c represent the input, forget and output gates and the cell state, respectively. For example, w_i is the weight on the linkage of the new input $x(t)$ to the input gate of a hidden neuron while w_{ih} is the weight on the linkage of the previous neuron output $h(t-1)$ to the input gate of a hidden neuron. σ represents the sigmoid function, $tanh$ represents the hyperbolic tangent function, and $*$ represents the Hadamard product.

At the individual pixel level, we constructed one-hidden-layer LSTM networks, due to the relatively small amount of data available (i.e., a total of 252 time steps). The numbers of neurons on the input and output layers depend on the numbers of input and output variables, respectively, and they are constant in each

experiment. Therefore, in this study, the network complexity is only affected by the number of hidden neurons, which is the only hyperparameter to tune during network validation (described in Section 3.2.4). The desired number of hidden neurons should allow a network not only to gain enough knowledge from a given data set but also be able to handle previously unobserved data (Dawson and Wilby, 2001; Müller and Guido, 2017b).

3.2.2 Wavelet coherence analysis

Wavelet transforms map time series into the time-frequency domain and help localizing intermittent periodicities. Continuous wavelet transform (CWT) is a common type of wavelet transforms useful for feature extraction (Grinsted et al., 2004). The CWT of a discrete time series x_{n_0} at the time step n and a specific time scale s is given by Eq. (3.7).

$$W(s,n) = \sum_{n_0=0}^{N-1} x_{n_0} \psi^*[(n_0 - n)\delta t / s] \tag{3.7}$$

where ψ is the mother wavelet, here using the Morlet wavelet, and ψ^* is the complex conjugate of ψ. δt is the time step of x_{n_0}, and N is the total number of δt in x_{n_0}. The power of the CWT is defined as $|W(s,n)|^2$ (Torrence and Compo, 1998). While CWT can effectively identify localized intermittent oscillations in the time-frequency domain, it is only applicable to a single time series.

Wavelet coherence analysis is a method that measures the cross-correlation of two time-dependent variables in the time-frequency domain, of which calculation (Eq. 3.8) is based on CWT. The results of wavelet coherence analysis are comparable with traditional correlation coefficients, ranging from 0 to 1.

$$Coherence(s,n) = \left|\langle s^{-1} W_{xy}(s,n)\rangle\right|^2 \Big/ \left(\langle s^{-1}|W_x(s,n)|^2\rangle \langle s^{-1}\left|W_y(s,n)\right|^2\rangle\right) \tag{3.8}$$

where the $\langle \cdot \rangle$ indicates smoothing in both time and scale. $W_x(s,n)$ and $W_y(s,n)$ are the CWT of the time series x_{n_0} and y_{n_0}. $W_{xy}(s,n)$ is the cross-wavelet spectrum of the time series x_{n_0} and y_{n_0} and equal to $W_x(s,n)W_y^*(s,n)$, where the (*) indicates the complex conjugate (Torrence and Webster, 1999; Grinsted et al., 2004).

The phase shift in the results of wavelet coherence analysis is calculated by:

$$\phi(s,n) = \tan^{-1}\left(\Im\{\langle s^{-1} W_{xy}(s,n)\rangle\} / \Re\{\langle s^{-1} W_{xy}(s,n)\rangle\}\right) \tag{3.9}$$

where, $\Im\{\cdot\}$ and $\Re\{\cdot\}$ signify the imaginary and real parts of a complex number, respectively (Torrence and Webster, 1999). $\phi(s,n) = 0$ means that the wavelets of two considered time series at the time step n and the time scale s are in phase. Detailed descriptions of wavelet coherence analysis and phase shift can be found in Torrence and Webster (1999), Grinsted et al. (2004) and Rahmati et al. (2020).

In this study, we conducted wavelet coherence analysis to derive the time-frequency correlation and phase shifts between regionally averaged wtd_a time series obtained from the TSMP-G2A data set and the LSTM network results. In this way, we expected to gain an understanding of the network performance in various experiments and thus help explain the contributions of different input variable combinations to the estimation of wtd_a over Europe in the time-frequency domain.

3.2.3 Study area and data set

We utilized the TSMP-G2A data set to evaluate the ability of the proposed LSTM networks to estimate wtd_a over Europe in different experiments. As aforementioned, the dataset contains daily averaged continuous simulation results over Europe from TSMP, which is a fully coupled atmosphere-land-surface-subsurface modeling system. The spatial resolution of the dataset is 0.11° (~ 12.5 km). For details regarding TSMP and the TSMP-G2A data set, the reader is referred to Shrestha et al. (2014), Gasper et al. (2014) and Furusho-Percot et al. (2019).

This study focused on eight different European regions, which are known as the PRUDENCE (Prediction of Regional scenarios and Uncertainties for Defining EuropeaN Climate change risks and Effects) regions (Christensen and Christensen, 2007): Scandinavia (SC), British Isles (BI), Mid-Europe (ME), Eastern Europe (EA), France (FR), Alps (AL), Iberian Peninsula (IP), and Mediterranean (MD). The individual PRUDENCE regions are characterized by different hydrometeorological regimes, potentially resulting in various responses of GW anomalies to other hydrometeorological variables, and constitute reference regions in climate science. Regionally averaged precipitation, evapotranspiration (ET) and soil moisture (θ) values calculated from the TSMP-G2A data set range from 776 mm (EA) to 1494 mm (AL), 283 mm (SC) to 518 mm (MD) and 0.29 m^3m^{-3} (IP) to 0.36 m^3m^{-3} (BI), respectively. Regionally averaged wtd values are from 2 m to 3 m, apart from AL (4.14 m), IP (6.62 m) and MD (6.48 m). Figure 3.2 presents the yearly averaged spatially distributed wtd values calculated from the TSMP-G2A data set from 01/1996 to 12/2016. Based on intervals of yearly averaged wtd, we categorized pixel values into three classes following Chapter 2, that is, C1 corresponding to 0.0 m < wtd ≤ 3.0 m, C2 corresponding to 3.0 m < wtd ≤ 10.0 m and C3 corresponding to wtd > 10.0 m. Most pixels on the European continent belong to C1 (colored in light blue), accounting for 52% to 75% of different PRUDENCE regions. Less than 20% of pixels in each PRUDENCE region belong to C2 (colored in orange). Less than 15% of pixels in each PRUDENCE region belong to C3 (colored in red), except for AL (24%), IP (30%) and MD (28%). In addition, there is also a heterogeneous pattern in the yearly averaged snow water equivalent (SWE) values derived from the TSMP-G2A data set (see Table 2.1). SC and AL have the largest regionally averaged SWE (> 60 mm) while the other regions have regionally averaged SWE < 10 mm.

The anomaly data were calculated from the TSMP-G2A data set for each calendar month and pixel individually for the period 01/1996 to 12/2016 to enable the spatial comparability and to account for the seasonality in the variables. Here, we only considered the data after the year 1996 to avoid the influence of spinup effects on the simulation results (Furusho-Percot et al., 2019). We used Eq. (3.10) to calculate pr_a,

ET_a, θ_a and wtd_a, where the climatological average and standard deviation values were derived from the training set (i.e., the data from 01/1996 to 12/2012, described in Section 3.2.4) to prevent future information from leaking into the training process. The rs_a is equal to wtd_a at pixels where $wtd \leq 0$ m. Regionally averaged pr_a, ET_a, θ_a and wtd_a time series for the *wtd* categories C1 to C3 in different PRUDENCE regions are presented in Figure B.1.

$$v_a = (v_m - v_{av})/v_{sd} \tag{3.10}$$

where v is the considered variable, such as *wtd*, v_m is the monthly data of v calculated from the TSMP-G2A data set, v_{av} is the climatological average of v_m (i.e., averages of v_m in January, February, …, December), and v_{sd} is the climatological standard deviation of v_m.

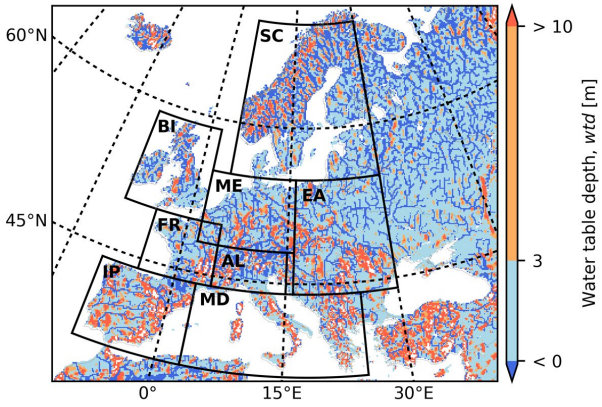

Figure 3.2: TSMP-G2A *wtd* [m] climatology over the European continent from 01/1996 to 12/2016. Areas bounded by the black polygons show PRUDENCE regions (i.e., SC - Scandinavia; BI - British Isles; ME - Mid-Europe; EA - Eastern Europe; FR - France; AL - Alps; IP - Iberian Peninsula; MD - Mediterranean).

The pr_a, ET_a, θ_a, rs_a and wtd_a are measures of different types of drought. Table 3.1 provides the definition of drought severity based on anomalies, following McKee et al. (1993).

Table 3.1: Definition of drought severity based on anomalies.

Drought severity	pr_a / θ_a	ET_a / rs_a / wtd_a
Extreme drought	≤ -2	≥ 2
Severe drought	-2 – -1.5	1.5 – 2
Moderate drought	-1.5 – -1	1 – 1.5
Minor drought	-1 – 0	0 – 1
No drought	> 0	< 0

Moreover, we utilized Eqs. (3.11) and (3.12) to calculate SWE_{scaled} from *SWE* which has data only available from 01/2003 to 12/2010. The obtained SWE_{scaled} data only has one value in the range from -1 to 1 at each pixel, so it is static.

$$SWE_0 = (SWE_y - SWE_{min})/(SWE_{max} - SWE_{min}) \tag{3.11}$$

$$SWE_{scaled} = SWE_0 * [1 - (-1)] + (-1) \tag{3.12}$$

where SWE_y is the yearly averaged *SWE* value from 2003 to 2010, SWE_{min} is the minimum value of SWE_y over the European continent, and SWE_{max} is the maximum value of SWE_y over the European continent.

3.2.4 Experimental design

In this study, we varied a number of input variables, in addition to pr_a used in the LSTM networks proposed in Chapter 2, to arrive at improved wtd_a estimates. The combinations of input variables used in different experiments are listed in Table 3.2. We selected the input variables based on their demonstrated relationship with wtd_a (Van Loon, 2015) and availability in the TSMP-G2A data set and common observational datasets. In E1, the LSTM networks used combinations of pr_a, ET_a and θ_a as input. The pr_a, ET_a and θ_a are drought-related hydrometeorological variables with spatiotemporally continuous data over Europe, which can be easily obtained from observations and reanalysis datasets, e.g., ERA5-Land. The pr_a and ET_a are measures of meteorological drought, while the θ_a shows the degree of agricultural drought. Except for the input variables involved in E1, the quality of wtd_a estimates can also be affected by *SWE*; Chapter 2 found that large *SWE* can degrade the network test performance. Moreover, in an unconfined aquifer, GW and surface water dynamics have a strong lateral connection, and thus, wtd_a at a pixel is also influenced by the change in water dynamics at neighboring pixels, especially for a pixel close to a river, due to the interaction between surface water and GW. Here, in addition to the optimal input variable combination determined in E1, we introduced a static input SWE_{scaled} and anomalies at adjacent pixels as input to the LSTM networks in E2, to explore potential improvement in the network performance. Especially, at pixels close to rivers, we investigated the impact of rs_a at adjacent pixels on the network performance (see E2.3).

Table 3.2: Combinations of input variables in different experiments.

Experiment		Combination of input variables
E1	E1.1	pr_a
	E1.2	ET_a
	E1.3	θ_a
	E1.4	pr_a, ET_a
	E1.5	pr_a, θ_a
	E1.6	ET_a, θ_a
	E1.7	pr_a, ET_a, θ_a
E2	E2.1	Optimal input variable combination in E1, SWE_{scaled}
	E2.2	Optimal input variable combination in E1 at the considered pixel and adjacent pixels
	E2.3	Optimal input variable combination in E1, rs_a at adjacent pixels

At the individual pixel level, we divided anomaly data into a training set (01/1996 - 12/2012, 204 time steps, about 80% of the total data), a validation set (01/2013 - 12/2014, 24 time steps, about 10% of the total data) and a test set (01/2015 – 12/2016, 24 time steps, about 10% of the total data) for network training, validation and testing, respectively. The static input SWE_{scaled} provided the same value at every time step.

The LSTM networks applied here have the same configuration of hyperparameters (listed in Table 3.3) as Chapter 2, except for the number of input neurons which depends on the number of input variables used in the different experiments. As described in Section 3.2.1, the number of hidden neurons has a significant impact on the network performance and is the only hyperparameter to tune here, ranging from 1 to 100.

Table 3.3: Hyperparameter setting of the applied LSTM networks.

Hyperparameter	Value or method
Number of input, hidden, and output layer(s)	(1, 1, 1)
Number of input, hidden and output neuron(s)	(number of input variables, 1-100, 1)
Initial weights and biases on all connections between neurons	U(-0.5, 0.5)*
Initial cell states of LSTM neurons	0
Optimizer and learning rate	RMSprop (Hinton et al., 2012), 0.001
Loss function	Mean square error (MSE)

* U(-0.5, 0.5) - uniform distribution bounded by -0.5 and 0.5.

Figure 3.3 illustrates the generic workflow used to construct the LSTM networks at the individual pixel level in the different experiments. The workflow started with the network training process during which we fitted the training set to the LSTM networks with 1 to 100 hidden neurons. An epoch is an iteration when the whole training set travels once through the network forward and backward. Weights and biases on all connections between neurons commenced from random values selected from a uniform distribution bounded by -0.5 and 0.5, and in each epoch, the networks automatically updated weights and biases based on the difference between the network output and the target data (i.e., TSMP-G2A wtd_a) calculated by the loss function (here mean square error, MSE). The technique of adjusting weights and biases is termed an optimizer (here RMSprop), and the rate at which an optimizer adjusts weights and biases is termed learning rate (here 0.001). In the interval of two consecutive epochs, the network validation process was run to check the performance of the trained network in each epoch on the validation set. The criteria for stopping the training process were 1) the number of epochs ≥ 50; and 2) the validation performance started to decrease. Based on the validation performance, we determined three optimal numbers of hidden neurons, which were often various for the LSTM networks constructed at different pixels. In the end, we applied the trained networks with the three optimal numbers of hidden neurons on the test set (i.e., unobserved data during training) and averaged the results from the three networks as the final result for evaluation. In this way, we compensated for the individual shortcomings of the selected networks to some extent and obtained improved final results. For details about the network setup, the reader is referred to Section 2.2.5.

We adopted RMSE and NSE scores as the metrics to assess the network performance, which show the goodness of fit of the LSTM networks in terms of the magnitude and variance of the error. They are

calculated by Eqs. (3.13) and (3.14), and the LSTM networks with good performance are expected to obtain low RMSEs and high NSE scores.

$$RMSE = \sqrt{\sum_{i=1}^{N}(y_t - y_e)^2 / N} \tag{3.13}$$

$$NSE = 1 - \sum_{i=1}^{N}(y_t - y_e)^2 / \sum_{i=1}^{N}(y_t - \bar{y_t})^2 \tag{3.14}$$

where y_t and $\bar{y_t}$ are the target value (i.e., TSMP-G2A wtd_a) and the average of the target values, respectively. The y_e is the estimated value (i.e., the network output, hereafter named LSTM wtd_a). N is the number of time steps in the given time series.

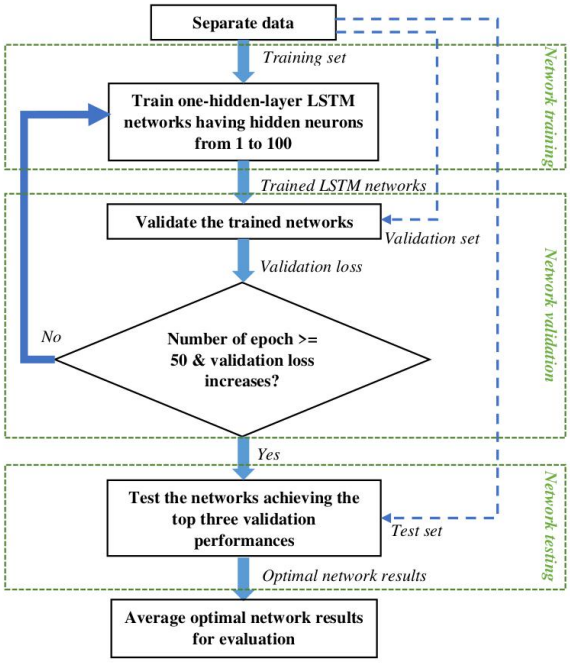

Figure 3.3: A generic workflow to construct the LSTM networks at individual pixels. The blue dashed lines with arrows represent data transmission paths.

To save computational resources, we constructed the LSTM networks locally at a limited number of pixels, which were randomly selected in each PRUDENCE region. Chapter 2 found that yearly averaged *wtd* considerably affected the test performance of the LSTM networks with pr_a as input, and for increasing *wtd*, the networks tended to behave poorly. Hence, during the analyses of the network results (presented in Section 3.3), we separated the selected pixels in each PRUDENCE region into the *wtd* categories C1 to C3 based on their *wtd* values.

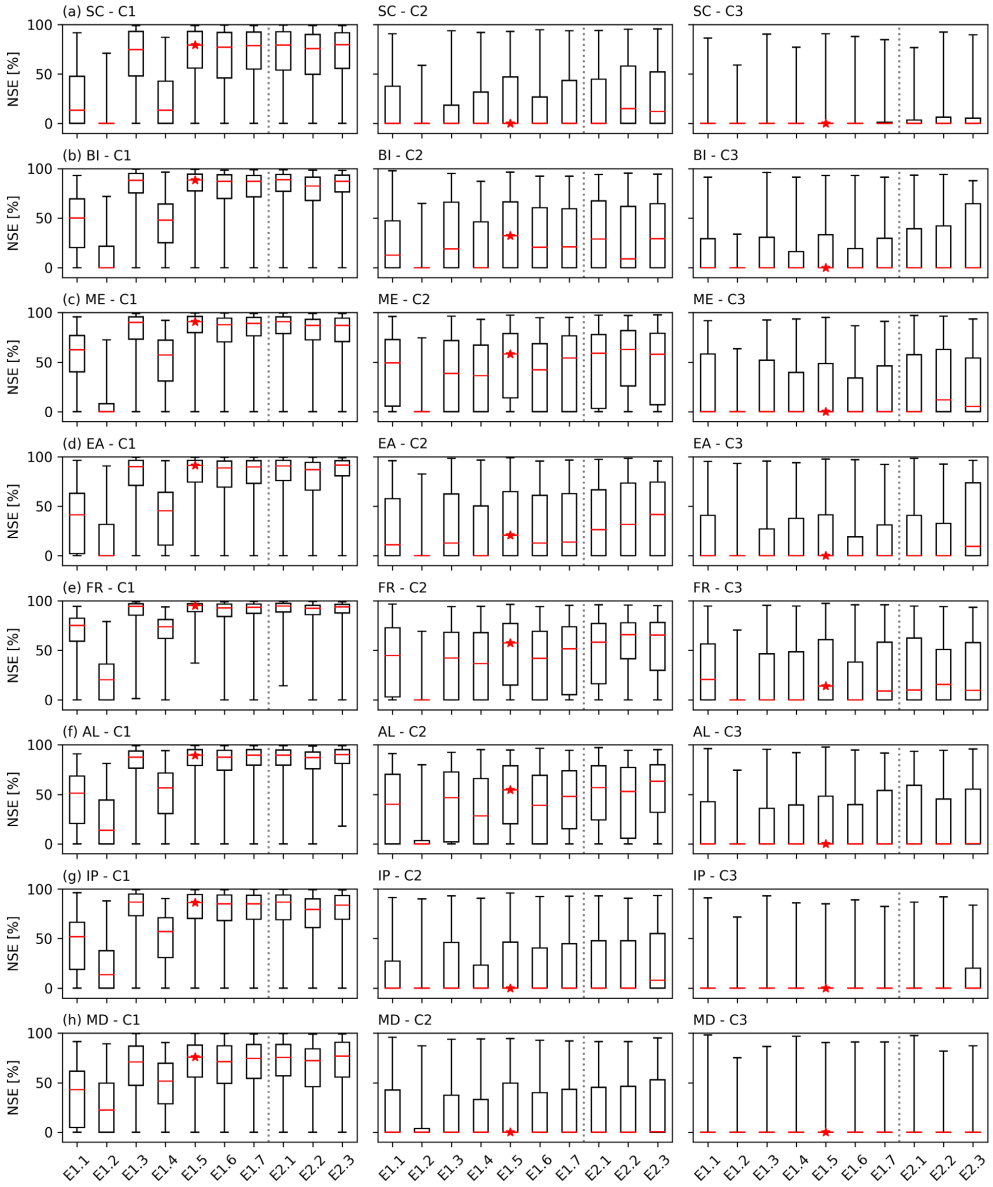

Figure 3.4: Box plots of test NSE scores achieved by the LSTM networks of E1 and E2: E1.1: pr_a; E1.2: ET_a; E1.3: θ_a; E1.4: pr_a and ET_a; E1.5: pr_a and θ_a; E1.6: ET_a and θ_a; E1.7: pr_a, ET_a and θ_a; E2.1: pr_a, θ_a and SWE_{scaled}; E2.2: pr_a, θ_a at the selected pixels and adjacent pixels; and E2.3: pr_a, θ_a at the selected pixels close to rivers and rs_a at the adjacent pixels. (a) – (h) show the comparison of the box plots at the selected pixels belonging to the *wtd* categories C1 to C3 in each PRUDENCE region. The box plots show the ranges of the NSE scores from the first quartile to the third quartile; the red lines indicate the medians of the NSE scores; and the upper and lower ends represent the maximum and minimum values of the NSE scores, respectively. The medians of the NSE scores obtained by the LSTM networks of E1.5 are marked with red stars. The box plots for E1 and E2 are separated by gray dotted lines.

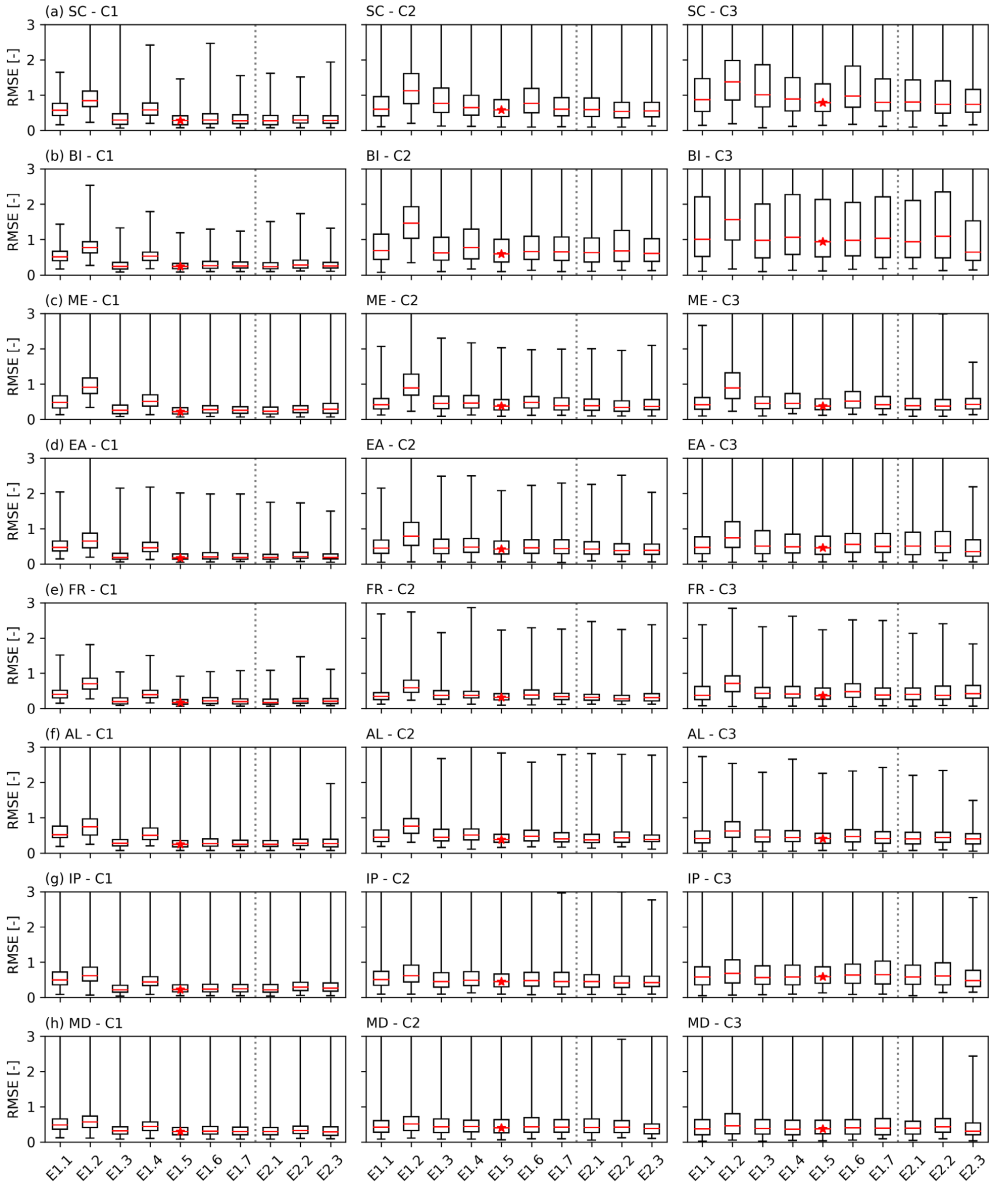

Figure 3.5: Box plots of test RMSEs achieved by the LSTM networks of E1 and E2. (a) – (h) show the comparison of the box plots at the selected pixels belonging to the *wtd* categories C1 to C3 in each PRUDENCE region. The labels have the same definitions as Figure 3.4, but for RMSEs.

3.3 Results

3.3.1 Test performance of the LSTM networks in the different experiments

We aimed to identify the LSTM networks with the best test performance in the designed experiments in handling previously unobserved data. Figures 3.4 and 3.5 illustrate the box plots of the test NSE scores and RMSEs achieved by the LSTM networks in the different experiments at the selected pixels belonging to the *wtd* categories C1 to C3 in each PRUDENCE region, respectively. For a better visualization, we set negative NSE scores to zero in Figure 3.4. The medians of the test NSE scores and RMSEs achieved in different experiments for C1 to C3 in each PRUDENCE region are listed in Table B.1.

The LSTM networks of E1.5 (marked with red stars in Figures 3.4 and 3.5) achieved the optimal test performance not only in E1 but also in all the designed experiments, which used pr_a and θ_a as input. They obtained the test scores as follows: median NSE of 76%-95% for C1, 0%-58% for C2, and 0%-14% for C3; and median RMSEs of 0.17-0.30 for C1, 0.32-0.60 for C2, and 0.36-0.94 for C3. The evaluation metrics were significantly improved compared to those obtained by the LSTM networks employed in Chapter 2, particularly for C1 (with a 20%-66% increase in median NSE and a 0.19-0.30 decrease in median RMSEs), which is the major *wtd* category in Europe. Over Europe, the LSTM networks of E1.5 showed good test performance with a test NSE score ≥ 50% at most selected pixels (Figure 3.6). In addition, Table 3.4 gives close to or over half of the selected pixels with a test NSE score ≥ 50% in the PRUDENCE regions for the LSTM networks of E1.5, constituting an increase of 8% to 22% compared to Chapter 2. The highly improved wtd_a estimates by including θ_a as input suggest that θ_a plays an important role in explaining GW dynamics over Europe. One possible reason for that is, compared with pr_a and ET_a, θ_a provides more information about subsurface hydrological processes, such as vegetation influence, soil heterogeneity and, thus, varying infiltration and recharge rates. Because θ_a and wtd_a are measures of agricultural drought and GW drought, respectively, the substantial contribution of θ_a to wtd_a also suggests the close connection between agricultural drought and GW drought over Europe.

In E2.1, no noticeable improvement was detected in the test NSE scores and RMSEs especially in SC and AL that have large *SWE* (see Figures 3.4 and 3.5) by adding SWE_{scaled} as additional input to the LSTM networks of E1.5, implying little additional contribution of snow accumulation to the estimation of wtd_a over Europe. This seems to be inconsistent with the conclusion of Chapter 2 that *SWE* strongly affects the quality of wtd_a estimates. However, the discrepancy can be explained by the fact that θ is partly replenished by snow water and thus θ_a already includes information about *SWE*.

In Figures 3.4 and 3.5, we found that only the test NSE scores and RMSEs for the *wtd* categories C2 and C3 (*wtd* > 3.0 m) were partially improved by adding anomalies available at neighboring pixels as input to the LSTM networks of E1.5 (i.e., E2.2 and E2.3). It indicates the lateral GW exchange mainly influences GW dynamics in deep aquifers. The improvement was small (< 0.1 for median RMSEs) for C2 and C3, and the network test performance was generally degraded for C1 to which more than half of the pixels over

Europe belong, and thus, the LSTM networks of E2.2 and E2.3 are expected to gain worse test performance compared to the LSTM networks of E1.5 at the European scale.

In general, Figures 3.4 and 3.5 show the decrease in network test performance with increasing average *wtd* (from C1 to C3), manifested in the smaller medians of test NSE scores and larger medians of test RMSEs, which is in good agreement with the finding in Chapter 2. Moreover, we also found only small contributions of ET_a to the estimation of wtd_a at the European scale compared to other drought-related input variables reflected in the worst test performance in E1.2 (ET_a).

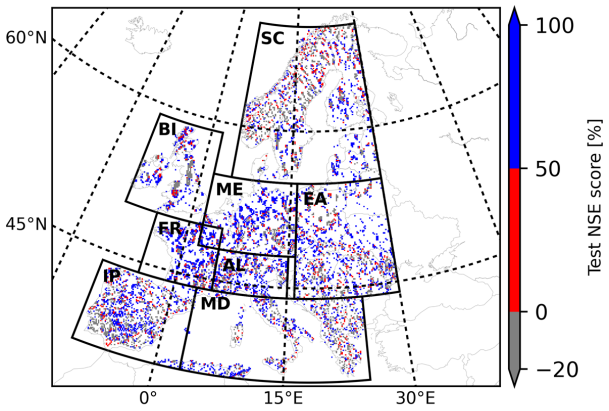

Figure 3.6: Map of test NSE scores achieved by the LSTM networks of E1.5 (pr_a and θ_a) in the PRUDENCE regions.

Table 3.4: Percentages of the selected pixels with a test NSE score ≥ 50% in the PRUDENCE regions [%] for the LSTM networks of E1.5.

SC	BI	ME	EA	FR	AL	IP	MD
37.90	50.10	64.49	50.60	65.74	61.99	44.50	44.90

3.3.2 European water table depth anomaly map in 2015 reproduced by the results of the optimal LSTM networks

In 2015, large parts of the European continent were affected by an extreme summer heatwave, causing severe drought (Van Loon et al., 2017). Here, based on the results of the optimal LSTM networks (E1.5: pr_a and θ_a), we reproduced the European wtd_a map derived from the TSMP-G2A data set for August 2015 (Figure 3.7), which was one of the driest months in 2015. The study month is in the testing period, and thus, the networks have not seen the data during training. Nevertheless, the reproduced map is in good agreement with the original map (Figure 3.7) concerning drought severity through a visual inspection. Both show the spatially heterogeneous distribution of wtd_a over Europe, including the strong GW anomalies (wtd_a ≥ 1.5) in ME, EA, and AL and the very wet conditions (wtd_a ≤ -1.5) in SC and BI, which is consistent with previous studies (e.g., Dong et al., 2016; Van Lanen et al., 2016; Van Loon et al., 2017). Moreover, the

optimal LSTM networks successfully detected ~41% of strong drought events ($wtd_a \geq 1.5$) and ~29% of extreme drought events ($wtd_a \geq 2$) at the European scale in August 2015, outperforming the original LSTM networks proposed in Chapter 2 (E1.1, Figure B.2b) with the hit rates only ~15% for strong drought events and ~3% for extreme drought events.

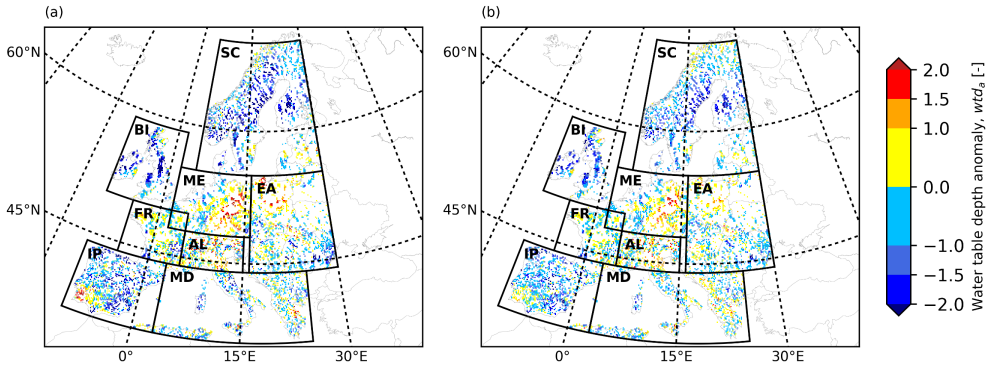

Figure 3.7: European wtd_a maps for August 2015 (i.e., in the testing period) derived from (a) the TSMP-G2A data set and (b) the results from the LSTM networks of E1.5 (pr_a and θ_a).

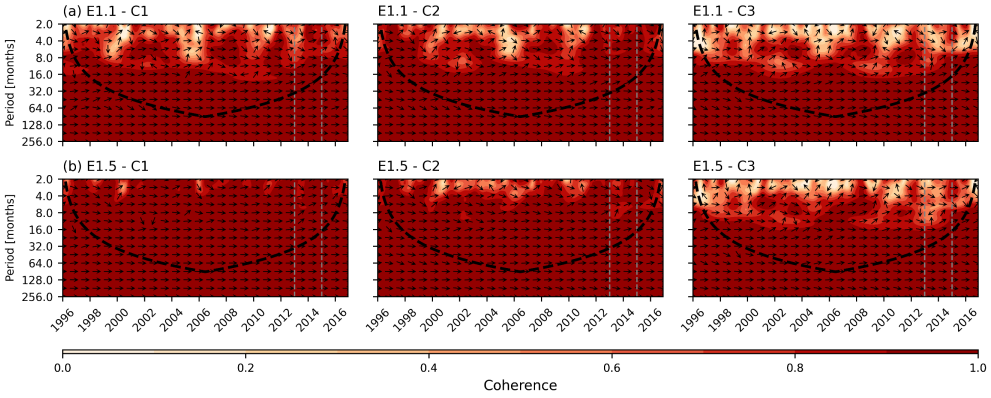

Figure 3.8: Results of wavelet coherence analysis on the regionally averaged wtd_a time series for the *wtd* categories C1 to C3 in ME, which were derived from the TSMP-G2A data set and the results of the LSTM networks of: (a) E1.1: pr_a; and (b) E1.5: pr_a and θ_a. The training, validation and testing periods are separated by gray dashed lines.

3.3.3 Wavelet coherence analysis on regionally averaged water table depth anomaly time series in ME

The significantly improved test performance of the LSTM networks of E1.5 is attributed to information contained in θ_a for estimating certain frequency components of wtd_a. Therefore, here we conducted wavelet coherence analysis on the regionally averaged wtd_a time series for the *wtd* categories C1 to C3 in ME (Figure 3.8), which were derived from the TSMP-G2A data set and the LSTM network results in E1.1 and

E1.5. We focused on the areas within the black dashed lines to eliminate edge effects and to gain a better understanding of the contribution of θ_a on the explanation of GW anomalies in the time-frequency domain.

By introducing θ_a as additional input, the coherence between the regionally averaged wtd_a time series was significant improved at periods between 2 and 8 months, especially for the *wtd* categories C1 and C2, revealing the larger contribution of θ_a to explain GW dynamics at the monthly and seasonal cycles compared to pr_a. Moreover, in Figure 3.8b, most phase shifts at periods between 2 and 8 months are zero for the *wtd* categories C1, suggesting almost no time lag between the estimated and target regionally averaged wtd_a time series in shallow aquifers of ME. The similar conclusions can also be drawn from the results of wavelet coherence analysis in other PRUDENCE regions, which are shown in Figures B.3-B.9.

3.4 Discussion

The scarcity of *wtd* observations complicates GW monitoring and requires alternative methods to quantify or predict wtd_a. The pr_a is the most common proxy of wtd_a, mainly due to the close connection between meteorological and GW droughts and the easy access to global precipitation data. This study, however, showed the limits of merely using pr_a and/or ET_a data to estimate wtd_a over Europe. Similar conclusions were also drawn by e.g., Kumar et al. (2016) and Uddameri et al. (2019), who compared the performance of SPI over extended time scales to quantify GW drought with the Standardized Groundwater level Index. One potential reason is that the occurrence of GW drought depends not only on the precipitation and temperature anomalies but also on the antecedent water storage (Van Lanen et al., 2016). Therefore, including ground-based information such as θ_a significantly improved the network results in the presented study. In addition, similar to precipitation, θ data is available from e.g., remotely sensed observations and reanalysis products, which removes the barriers to using θ_a as input for estimating wtd_a in real world applications.

The 2015 European summer heatwave started in June and resulted in peak temperatures in early July (Dong et al., 2016). Because most aquifers in Europe are shallow (with simulated $wtd \leq 3$m), a rapid GW response was noticed, and the GW drought was already severe in several parts of Europe in August 2015. Although the impact of the GW drought continued until the end of 2015 (not shown here), its affected area was much smaller than the meteorological drought in the same year (Dong et al., 2016; Van Lanen et al., 2016). This indicates that not all meteorological droughts will evolve into GW droughts, and thus simply considering precipitation information is not enough to quantify GW drought, which is consistent with our previous observation. Moreover, the locations with GW drought coincided well with those with vegetation stress presented in Van Lanen et al. (2016), which also confirms the usefulness of θ information in the estimation of GW anomalies.

The wavelet coherence analysis helped to explain the added value of θ_a as input in the estimation of wtd_a from the time-frequency perspective. Considering θ_a consistently increased the coherence between

regionally averaged TSMP-G2A and LSTM wtd_a time series at periods between 2 and 8 months. This reflects the systematic contribution of θ_a to the explanation of GW anomalies.

Temporal Convolutional Networks (Yan et al., 2020) and Transformers (Vaswani et al., 2017) may constitute alternatives to the LSTM networks proposed here, which have been shown to outperform LSTM networks in handling long time series. The estimation of GW anomalies in deep aquifers may benefit from the application of these methods. Yet, in the simulations, most aquifers in Europe have yearly averaged *wtd* value ≤ 3m, in which the response of wtd_a to hydrometeorological variables is expected to be relatively fast. Therefore, the improvement achieved by these methods may be not significant compared to LSTM networks.

3.5 Summary and conclusions

In this study, we conducted several experiments to investigate the impacts of additional input variable combinations in the LSTM networks proposed in Chapter 2 to improve monthly wtd_a estimates at individual pixels in eight hydrometeorologically different regions over Europe (i.e., PRUDENCE regions). Except for the original input variable pr_a, we introduced ET_a, θ_a, SWE_{scaled}, and anomalies at adjacent pixels (e.g., rs_a, see Table 3.2) as optional input variables to the LSTM networks. All assessments were based on anomalies derived from the TSMP-G2A data set, which contains daily integrated hydrologic simulation results over the European continent.

Because NSE scores and RMSEs only provide limited information on the network performance, we also applied wavelet coherence analysis to investigate the contribution of the input variable(s) to explain GW anomalies over Europe in the time-frequency domain.

The optimal LSTM networks were found with pr_a and θ_a as input. Considering θ_a strongly improved the network test performance particularly in the areas with *wtd* ≤ 3 m (i.e., the major *wtd* category), suggesting that θ_a plays a significant role in the estimation of wtd_a over Europe. Because θ_a is related to agricultural drought and wtd_a shows the degree of GW drought, we conclude that a strong link exists between agricultural drought and GW drought on the European continent. The proposed LSTM networks can generate good results in shallow aquifers but may fail in deep aquifers. Therefore, one should be careful to use such networks in areas with large *wtd*.

We recognize that the network performance was limited by the relatively small amount of available training data, the simplified tuning of hyperparameters, and the application of simulation data (i.e., the TSMP-G2A data set) for evaluation. The biases of the TSMP-G2A data set mainly come from uncalibrated parameters and neglect of human impacts (Furusho-Percot et al., 2019). Nevertheless, due to the good agreement of the TSMP-G2A data set with hydrological observational datasets, we argue that the methodology is useful in determining the optimal input variable combination for the estimation of wtd_a over Europe. Our study highlights the benefit in combining θ information with precipitation information to estimate wtd_a over Europe. The input of the LSTM networks should have valid values continuous in time. In the future, the indirect

method based on the optimal LSTM networks may be transferred to real-time monitoring of GW drought at the European scale using remotely sensed surface θ observations from e.g., the SMOS (Kerr et al., 2010) and SMAP (Entekhabi et al., 2010) missions and precipitation observations from e.g., the GPM (Hou et al., 2014) and TRMM (Huffman et al., 2007) missions. Similarly, also some numerical weather prediction models provide the necessary variables, e.g., the ECMWF Integrated Forecasting System with assimilated ASCAT (MetOp Advanced SCATerometer) θ data (Aires et al., 2021).

Chapter 4 Advancing AI-based pan-European groundwater monitoring

The content of this chapter will be submitted for publication.

Ma, Y., Montzka, C., Naz, B. and Kollet, S.: Advancing AI-based pan-European groundwater monitoring, in preparation.

Abstract

The main challenge of pan-European groundwater (GW) monitoring is the lack of collated water table depth (*wtd*) observations. Combining Long Short-Term Memory (LSTM) networks and transfer learning (TL), we propose an AI-based methodology LSTM-TL to produce reliable *wtd* anomaly (wtd_a) estimates at the European scale in the absence of consistent *wtd* observational data sets. With substantially reduced computational cost, LSTM-TL obtained wtd_a estimates in good agreement with *in-situ* wtd_a measurements from approximately 2,600 European GW monitoring wells, showing $r \geq 0.5$, RMSE ≤ 1.0 and KGE ≥ 0.3 at about or more than half of the pixels. Based on the reconstructed long-term European monthly wtd_a data from the early 1980s to the near present, we provided the first estimate of seasonal wtd_a trends in different parts of Europe, which facilitates the understanding of historical GW dynamics in Europe.

4.1 Introduction

Groundwater (GW) is the dominant source of drinking water in Europe, with about 75% of European Union residents relying on GW for their water supply, and recognized as an important contributor to industry (e.g., cooling water) and agriculture (irrigation) (European Commission, 2008). In addition, GW plays a critical role in sustaining surface water bodies, adapting to climate variability and supporting biodiversity (van der Gun, 2020). In recent years, however, Europe has experienced several severe summer droughts and heat waves (e.g., Fink et al., 2004; Stahl et al., 2016; Van Lanen et al., 2016; Bastos et al., 2020; Boergens et al., 2020), affecting all compartments of the water cycle. Droughts often affect large areas across national boundaries (Brauns et al., 2020). High ambient atmospheric temperature, continuous precipitation deficits, and large evapotranspiration losses translate into delayed and attenuated soil moisture (θ) deficits, which subsequently reduce GW recharge and increase water table depths (*wtd*) and ultimately, cause low streamflow or dried-up rivers (Tallaksen et al., 2009; Van Loon, 2015; Hellwig et al., 2020). More frequent and severe droughts anticipated in future changing climate will exacerbate the vulnerability of European GW systems, emphasizing the necessity of GW monitoring in European GW management (Guerreiro et al., 2018; Ault, 2020).

The major challenge for pan-European GW monitoring is the scarcity of collated *wtd* observations, which is attributed to the large spatial gaps and temporal inconsistency in the *wtd* measurement network (Brauns et al., 2020). Sparse *wtd* measurements pose difficulties in understanding GW dynamics at the continental scale and limit insight into extreme events (e.g., droughts) and their impacts on GW over Europe. To overcome the challenge, standardized meteorological drought indices over extended time scales and terrestrial water storage anomalies derived from the Gravity Recovery and Climate Experiment (GRACE) satellite observations have been widely used to predict and quantify GW anomalies (e.g., Bloomfield and Marchant, 2013; Kumar et al., 2016; Van Loon et al., 2017; Boergens et al., 2020). However, their reliability is often questioned, especially in small watersheds, mainly due to the unrealistic assumption about the linear translation of changes in meteorological signals (e.g., precipitation) into GW anomalies or the coarse spatial resolution (0.5°, about 55 km) of the GRACE observations (Kumar et al., 2016; Van Loon et al., 2017). Physically-based numerical models enable a realistic representation of processes in the terrestrial water and energy cycles. While they may provide more accurate GW anomaly estimates than the above methods, the models require extensive physical background knowledge and become too time and computationally demanding to be applied for high-resolution long-term simulations at larger scales (Wunsch et al., 2018; Hauswirth et al., 2021). Therefore, in this study, a methodology is proposed that offers reliable high-resolution long-term estimates of *wtd* anomalies (wtd_a) at the European scale with low computational cost, which allows effective and efficient GW monitoring over Europe.

With the emergence of deep learning (DL), increase in computational power, and availability of large datasets, the application of artificial intelligence (AI) has again been attracting considerable attention across scientific disciplines, including hydrological sciences. AI can produce hydrological predictions as accurately

as physically-based numerical models with much less background knowledge and computational cost (Govindaraju, 2000; Shen, 2018). As a subset of AI, DL techniques such as Long Short-Term Memory (LSTM) networks have achieved great success in predicting GW changes in previous studies (e.g., J. Zhang et al., 2018; Hauswirth et al., 2021; Vu et al., 2021). Specifically, LSTM networks are able to detect long-short-term dependencies between GW anomalies and input hydrometeorological forcings without explicitly defining the time lag, thereby simplifying the data preprocessing process (Hochreiter and Schmidhuber, 1997; Gers et al., 2000). In the field of hydrology, such DL techniques are generally trained using a supervised training algorithm with a supplementary teacher signal (here wtd_a observations, $wtd_{a,o}$) to guide the training process, as shown in Figure 4.1a. As a result, it is challenging to implement DL techniques to estimate wtd_a without sufficient $wtd_{a,o}$ used for training.

Transfer learning (TL), which is a machine learning technique (belonging to the broad field of AI) that applies the knowledge gained in a data-rich domain into a related data-scarce domain, can address the issue of sparse wtd_a observations to train DL techniques (Tan et al., 2018). The increased knowledge often significantly improves network performance in the data-scarce domain and/or accelerates computing progress (Goodfellow et al, 2017c), and thus, TL has become very popular in image classification (e.g., Quattoni et al., 2008; Lu et al., 2015; Lee et al., 2017) and natural language processing tasks (e.g., Lu et al., 2015). A recent study (Ma et al., 2021) transferring hydrological parameters across continents has also demonstrated the usefulness of TL in improving hydrological predictions. In contrast with sparse *in-situ* $wtd_{a,o}$ (the data-scarce domain), anomalies derived from modeling results, such as modeled wtd_a ($wtd_{a,m}$), are spatiotemporally continuous throughout the European continent and can be regarded as a data-rich domain. Modeling results from physically-based numerical models may contain the realistic relationship between wtd_a and input hydrometeorological variables, and thus, are used for knowledge transfer.

Here, we propose an AI-based methodology that integrates LSTM networks with TL, LSTM-TL, to estimate monthly wtd_a at the continental scale in order to facilitate pan-European GW monitoring. The methodology (illustrated in Figure 4.1b) uses monthly precipitation and θ anomalies (pr_a and θ_a) as input, which have observations and modeling results available at large scales, and relies on the close connection between GW and other compartments in the water cycle to produce wtd_a estimates. Firstly, the LSTM networks are trained at individual pixels on published modeling results (Furusho-Percot et al., 2019) from an uncalibrated fully coupled atmospheric-land-surface-groundwater modeling system (i.e., the Terrestrial Systems Modeling Platform, TSMP). Secondly, without additional training, the LSTM networks trained in the previous step are utilized to estimate wtd_a based on pr_a and θ_a from common observational datasets ($pr_{a,o}$ and $\theta_{a,o}$), thereby transferring the modeled input-output relationship (f_m) to the observation-based estimation. The implementation of LSTM-TL is based on two assumptions, that is, (i) f_m agrees well with the observed input-output relationship (f_o); and (ii) the internal LSTM networks successfully capture f_m during training, and the learned relationship is labeled as f_{lstm}.

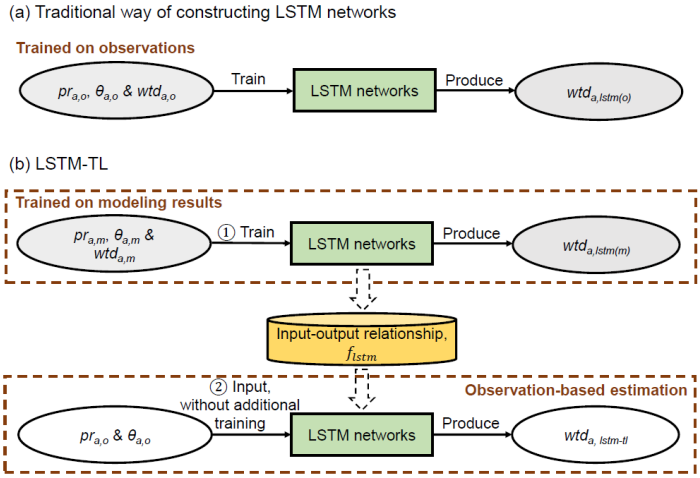

Figure 4.1: Comparison between the traditional way of constructing LSTM networks and LSTM-TL. (a) Tranditional way of constructing LSTM networks. LSTM networks are trained on observations, and their output is labeled as $wtd_{a,lstm(o)}$. (b) Schematic of LSTM-TL workflow to estimate wtd_a at individual pixels over Europe. Firstly, LSTM networks are trained on modeling results, and the output in this step is labeled as $wtd_{a,lstm(m)}$. Secondly, without additional training, the trained LSTM networks in the previous step are utilized to produce wtd_a estimates based on $pr_{a,o}$ and $\theta_{a,o}$, and the final output of LSTM-TL is labeled as $wtd_{a,lstm\text{-}tl}$. In this way, the modeled relationship between wtd_a and input hydrometeorological variables learned by the LSTM networks (i.e., f_{lstm}) is transferred to the observation-based estimation for wtd_a.

LSTM-TL has three salient characteristics. First, the methodology is independent of $wtd_{a,o}$ to estimate wtd_a, which allows its usage in large regions even without *wtd* observations. Second, the methodology can produce longer-term wtd_a estimates than modeling data used for training, and the time length of the obtained wtd_a estimates depends on input $pr_{a,o}$ and $\theta_{a,o}$. This is useful for reconstructing historical wtd_a and predicting future wtd_a at the continental scale. Third, once the internal LSTM networks are successfully trained, the methodology can be used without additional training, and thus, requires small computational cost to generate new wtd_a estimates.

In this work, we evaluated the obtained wtd_a estimates by LSTM-TL ($wtd_{a,lstm\text{-}tl}$) with collated *in-situ* $wtd_{a,o}$ from approximately 2,600 European GW monitoring wells (Figure 4.2) in order to explore the reliability of LSTM-TL. Using $pr_{a,o}$ and $\theta_{a,o}$ from various observational datasets as input to LSTM-TL, we reconstructed monthly wtd_a data at the European scale from the early 1980s to the near present and investigated GW changes over Europe in recent drought years. Finally, seasonal wtd_a trends were derived for different European regions from the reconstructed data. To the best of our knowledge, these trends are shown for the first time.

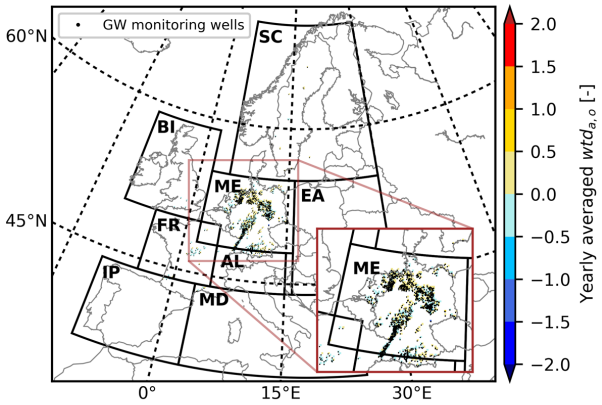

Figure 4.2: Yearly averaged $wtd_{a,o}$ [-] over the European continent from 01/1996 to 12/2016. Black circles mark the locations of 2,604 GW monitoring wells, at which *wtd* observations were collected. Areas bounded by black polygons show eight PRUDENCE regions (i.e., SC - Scandinavia; BI - British Isles; ME - Mid-Europe; EA - Eastern Europe; FR - France; AL - Alps; IP - Iberian Peninsula; MD - Mediterranean).

4.2 Methods

4.2.1 LSTM-TL

The study proposes LSTM-TL to estimate wtd_a at individual pixels over Europe in the absence of collated *wtd* measurements, thereby mitigating the negative impact of sparse measurements on European GW management. The methodology combines LSTM networks and TL. LSTM networks are utilized to capture the time-varying and time-lagged relationship between wtd_a and input hydrometeorological variables, while TL is applied to transfer the modeled input-output relationship (i.e., f_m) to the observation-based estimation of wtd_a. Here we select pr_a and θ_a as input, due to the large-scale availability of remotely sensed observations and reanalysis products. At pixels without $\theta_{a,m}$ information, pr_a is the only input variable. Figure 4.1b illustrates the workflow of LSTM-TL, encompassing two steps, i.e., training the internal LSTM networks on modeling results and estimating wtd_a based on $pr_{a,o}$, $\theta_{a,o}$ (observations) and f_{lstm}(the relationship learned by the internal LSTM networks, assuming $f_{lstm} \approx f_m$). The reader is referred to Chapters 2 and 3 for detailed information on the training, validation and testing processes of the internal LSTM networks. Once the internal LSTM networks are successfully trained, LSTM-TL can be repeatedly employed to generate new wtd_a data without additional training, which further saves computational resources. The time length of the obtained wtd_a estimates ($wtd_{a,lstm\text{-}tl}$) is consistent with input $pr_{a,o}$ and $\theta_{a,o}$.

Using 192 Intel Xeon 24-core AVX512 CPUs in parallel, it took LSTM-TL about 32 hours to complete the training of the internal LSTM networks at individual pixels over Europe. After successfully training of the LSTM networks, LSTM-TL generated $wtd_{a,lstm\text{-}tl}$ at the European scale on two 24-core CPUs in about two

hours. However, to produce wtd_a estimates at similar temporal and spatial scales, it usually requires physically-based numerical models running on more CPUs for weeks or even months.

4.2.2 Datasets

All anomalies here are relative to the 1996-2012 period (i.e., the training period of the LSTM networks in LSTM-TL), to avoid future information from leaking into the LSTM networks during training. The wtd_a was computed by Eq. (4.1) for each calendar month and pixel individually to ensure the spatial comparability and to account for the seasonality. The calculation of pr_a and θ_a is similar to wtd_a.

$$wtd_a = (wtd_m - wtd_{av})/wtd_{sd}, \tag{4.1}$$

where wtd_m is the monthly data of *wtd*, wtd_{av} is the climatological average of wtd_m (i.e., averages of wtd_m in January, February, …, December), and wtd_{sd} is the climatological standard deviation of wtd_m.

The monthly $pr_{a,m}$, $\theta_{a,m}$ and $wtd_{a,m}$ were derived from published modeling results of TSMP from 1996 to 2016, with a spatial resolution of 0.11° (Furusho-Percot et al., 2019). The current version (v1.1) of TSMP constitutes the numerical weather prediction model of COnsortium for Small-scale MOdeling (COSMO), v5.01, the National Center for Atmospheric Research (NCAR) Community Land Model (CLM), v3.5, and the 3D surface–subsurface hydrologic model ParFlow, v3.2, which are externally coupled by the Ocean Atmosphere Sea Ice Soil Model Coupling Toolkit (OASIS3–MCT) coupler (Gasper et al., 2014; Shrestha et al., 2014). Therefore, it closes the terrestrial water and energy cycle from GW via the land surface to the top of the atmosphere (Keune et al, 2016), which is unique.

The monthly $pr_{a,o}$ were derived from the ERA5 bias-corrected “Rainfall flux” and “Snowfall flux” data, the COSMO-REA6 “TOT_PRECIP” data and the ERA5-Land “Total precipitation” data at the individual pixel level, and the monthly $\theta_{a,o}$ were obtained from the ERA5-Land “Volumetric soil water layer 1” data (i.e., volume of soil water at 0-7 cm below the land surface) and the GLEAM v3.5a “SMsurf” at the individual pixel level. The ERA5 bias-corrected dataset is a global-scale hourly reanalysis product with a spatial resolution of 0.5°, ranging from 1979 to 2019 (Muñoz Sabater, 2021a). The ERA5 bias-corrected precipitation data (“Rainfall flux” + “Snowfall flux”) at hourly time steps have a mean absolute error < 0.15 mm h^{-1} across all 13 FLUXNET2015 sites, superior to its previous version (i.e., the WFDEI dataset, “WATCH Forcing Data methodology applied to ERA-Interim reanalysis data”) (Cucchi et al., 2020). The COSMO-REA6 dataset is a Europe-scale hourly reanalysis product with a spatial resolution of 0.055°, ranging from 1995 to August 2019. The COSMO-REA6 “TOT_PRECIP” data reproduces the diurnal cycle of precipitation intensities at > 1,000 weather stations across Germany for summer and winter. The probability distributions of 3 h aggregated precipitation sums from the COSMO-REA6 data and *in-situ* measurements agree well, e.g., in terms of the integrated quadratic distance (Bollmeyer et al., 2015). The ERA5-Land dataset is a global-scale hourly reanalysis product including θ with a spatial resolution of 0.1°, ranging from 1981 to May 2021 (Muñoz Sabater, 2021b). The GLEAM v3.5a dataset is a global-scale daily θ reanalysis product with a spatial

resolution of 0.25°, ranging from 1980 to 2020 (Miralles et al.,2011; Martens et al., 2017). The ERA5-Land and GLEAM *θ* data both agree well with *in-situ* measurements from > 800 sensors located primarily in the USA and Europe demonstrated in an intercomparison of 18 *θ* products (Beck et al., 2021). Before calculating anomalies, all input observational datasets were re-gridded to 0.11° × 0.11° via the first-order conservative interpolation method (Jones, 1999).

The monthly $wtd_{a,o}$ were derived from consecutive monthly *wtd* measurements at 2,604 GW monitoring wells (Figure 4.2) in Europe from 1996 to 2016. The *wtd* measurements were obtained either from web services or by request from governmental authorities; detailed information is provided in Table C.1. The 0.11 ° gridded $wtd_{a,o}$ data used for evaluation were estimated by averaging $wtd_{a,o}$ from all the wells that lie within the same 0.11° pixels. The number of wells within each 0.11° pixel varies from 1 to 49, with a median value of 2 (Figure C.1). The pixels are unevenly distributed in eight hydrometeorologically different regions over Europe (named PRUDENCE regions) (Christensen and Christensen, 2007), and their numbers are provided in Table C.2. The yearly averaged $wtd_{a,o}$ generally ranged from -0.5 to 0.5 at individual pixels for the 1996-2016 period, as illustrated in Figure 4.2.

4.2.3 Evaluation metrics

In this study, we used the Pearson correlation coefficient (r), root mean square error (RMSE), Nash-Sutcliffe efficiency (NSE) and Kling-Gupta efficiency (KGE) as evaluation metrics. The equations to calculate the scores of $wtd_{a,lstm-tl}$ based on $wtd_{a,o}$ are shown below. The scores of other wtd_a estimates were computed in a similar way. The ideal r, RMSE, NSE and KGE of wtd_a estimates are one, zero, one and one, respectively.

$$r = \frac{\sum(wtd_{a,lstm-tl} - \overline{wtd_{a,lstm-tl}})(wtd_{a,o} - \overline{wtd_{a,o}})}{\sqrt{\sum(wtd_{a,lstm-tl} - \overline{wtd_{a,lstm-tl}})^2 \sum(wtd_{a,o} - \overline{wtd_{a,o}})^2}} \tag{4.2}$$

$$RMSE = \sqrt{\frac{\sum(wtd_{a,o} - wtd_{a,lstm-tl})^2}{N}} \tag{4.3}$$

$$NSE = 1 - \frac{\sum(wtd_{a,o} - wtd_{a,lstm-tl})^2}{\sum(wtd_{a,o} - \overline{wtd_{a,o}})^2} \tag{4.4}$$

$$KGE = 1 - \sqrt{(r-1)^2 + (\frac{\sigma_{wtd_{a,lstm-tl}}}{\sigma_{wtd_{a,o}}} - 1)^2 + (\frac{\overline{wtd_{a,lstm-tl}}}{\overline{wtd_{a,o}}} - 1)^2} \tag{4.5}$$

where $\overline{wtd_{a,lstm-tl}}$ and $\sigma_{wtd_{a,lstm-tl}}$ are the mean and standard deviation of $wtd_{a,lstm-tl}$, respectively. $\overline{wtd_{a,o}}$ and $\sigma_{wtd_{a,o}}$ are the mean and standard deviation of $wtd_{a,o}$, respectively. N is the number of time steps in the $wtd_{a,lstm-tl}$ and $wtd_{a,o}$ time series.

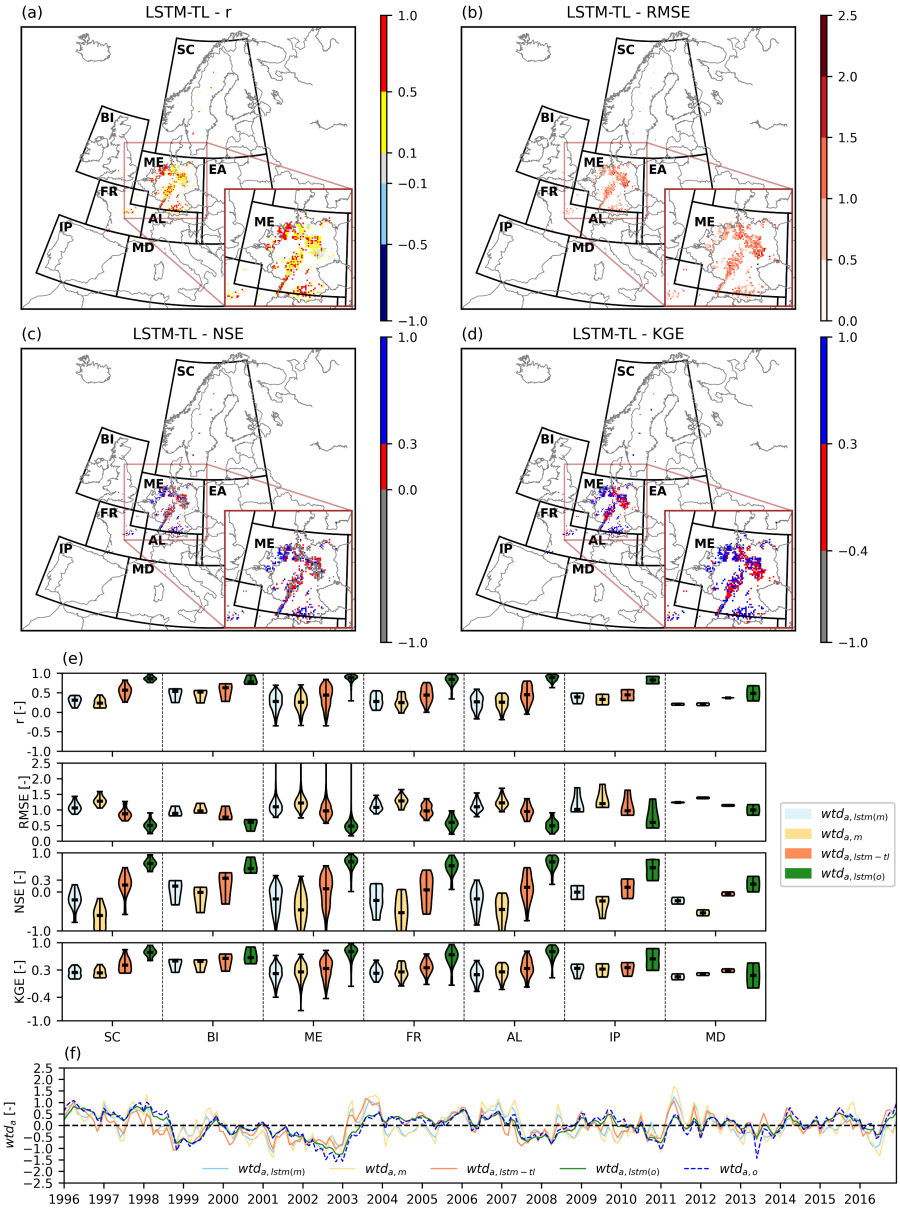

Figure 4.3: Performance of LSTM-TL in estimating wtd_a at pixels with $wtd_{a,o}$. (a)-(d) Maps of r, RMSE, NSE, KGE obtained by LSTM-TL at pixels with $wtd_{a,o}$ in Europe for the 1996-2016 period. (e) Violin plots show r, RMSE, NSE, KGE of $wtd_{a,lstm(m)}$, $wtd_{a,m}$, $wtd_{a,lstm-tl}$ and $wtd_{a,lstm(o)}$ in seven PRUDENCE regions for the 1996-2016 period. The violin plots reveal the distribution of scores in each region, and the bold line in each violin plot indicates the regional median of scores. (f) Regionally averaged $wtd_{a,lstm(m)}$, $wtd_{a,m}$, $wtd_{a,lstm-tl}$, $wtd_{a,lstm(o)}$ and $wtd_{a,o}$ time series in ME for the 1996-2016 period. The black dashed line indicates $wtd_a = 0$. A long time period with $wtd_a > 0$ corresponds to a dry event while a long time period with $wtd_a < 0$ corresponds to a wet event.

4.3 Results

4.3.1 LSTM-TL performance

LSTM-TL produced $wtd_{a,lstm\text{-}tl}$ at the European scale with a spatial resolution of 0.11° (about 12.5 km) from 01/1996 to 12/2016 based on anomalies from modeling results of TSMP and averaged $pr_{a,o}$ and $\theta_{a,o}$ from observational datasets (i.e., ERA5 bias-corrected, COSMO-REA6, ERA5-Land and GLEAM datasets). To assess the performance of LSTM-TL, we evaluated the obtained $wtd_{a,lstm\text{-}tl}$ with 0.11° gridded $wtd_{a,o}$ data, which were esimated from consecutive monthly $wtd_{a,o}$ data at around 2,600 European GW monitoring wells (Figure 4.2) from 01/1996 to 12/2016. The evaluation metrics used here are r, RMSE, NSE and KGE. The r shows the linear correlation between wtd_a estimates and $wtd_{a,o}$. RMSE and NSE reflect the goodness of fit of LSTM-TL in terms of the magnitude and variance of the error. Since anomalies were investigated here, the obtained RMSEs are unbiased. KGE is a simliar metric to NSE, which combines the three components of NSE (correlation, variability bias and mean bias) in a more balanced way (Knoben et al., 2019). The long-term means of wtd_a and $wtd_{a,o}$ are close to zero (per definition) at individual pixels, resulting in a high mean bias and ultimately a low and negative KGE, which is flawed. To avoid this, we set the mean bias term in the KGE to zero in this study. For pixels with $wtd_{a,o}$, LSTM-TL achieved r ≥ 0.5, RMSE ≤ 1.0, NSE ≥ 0.3 (NSE ≥ 0.0) and KGE ≥ 0.3 at 42%, 56%, 28% (57%) and 57% of the pixels for the 1996-2016 period, respectively (Figures 4.3a-d), showing reliable performance in estimating wtd_a over Europe. Regional differences were observed in the scores (Figure 4.3e). LSTM-TL always obtained the best scores in British Isles (BI), with a median r of 0.63, a median RMSE of 0.77, a median NSE of 0.35, and a median KGE of 0.60, and the worst scores in Mediterranean (MD), with a median r of 0.37, a median RMSE of 1.15, a median NSE of -0.05, and a median KGE of 0.29. We focused on the medians of the scores to reduce the influence of outliers on the overall performance evaluation.

We compared $wtd_{a,lstm\text{-}tl}$ with $wtd_{a,lstm(m)}$ (outputs of the LSTM networks using modeling results from TSMP as input), $wtd_{a,m}$ and $wtd_{a,lstm(o)}$ (outputs of the LSTM networks trained on observations) in seven PRUDENCE regions in terms of r, RMSE, NSE and KGE (Figure 4.3e), which were derived based on $wtd_{a,o}$. The medians of the scores are provided in Table C.3. Eastern Europe (EA) was not considered, which has no pixel with $wtd_{a,o}$. In general, the scores of $wtd_{a,lstm\text{-}tl}$ were improved compared to $wtd_{a,lstm(m)}$ and $wtd_{a,m}$. The largest improvement was found in SC, with an increase > 0.25 in median r and a decrease > 0.15 in median RMSE. The superiority of LSTM-TL in estimating wtd_a can be attributed to the correction of input data with $pr_{a,o}$ and $\theta_{a,o}$. The improved scores of $wtd_{a,lstm\text{-}tl}$ indirectly corroborates the realism of TSMP simulations in terms of hydrometeorological dynamics, which is consistent with the conclusions of Furusho-Percot et al. (2019) and Hartick et al. (2021). Similar distributions of the scores of $wtd_{a,lstm(m)}$ and $wtd_{a,m}$ displayed in the violin plots (Figure 4.3e) demonstrate the ability of the LSTM networks to capture f_m. The $wtd_{a,lstm(o)}$ is the only estimate that involved $wtd_{a,o}$ in its production, and thus, not surprisingly, received the best scores in almost all regions. Differences existed between the scores of $wtd_{a,lstm\text{-}tl}$ and $wtd_{a,lstm(o)}$ but

decreased for the period 2015-2016 (i.e., the test period when the LSTM networks were applied to previously unobserved data in the training, Figure C.2).

Mid-Europe (ME, zoomed in Figures 4.2 and 4.3a-d) has the most pixels with $wtd_{a,o}$ over Europe, i.e., 592 out of 681. Both regionally averaged $wtd_{a,lstm\text{-}tl}$ and $wtd_{a,lstm(o)}$ had good agreement with $wtd_{a,o}$ in ME (Figure 4.3f). The KGE of the regionally averaged $wtd_{a,lstm\text{-}tl}$ is 0.73, only 0.10 lower than that of the regionally averaged $wtd_{a,lstm(o)}$. This lends confidence in applying LSTM-TL in areas with limited or without *wtd* observations. In addition, the regionally averaged $wtd_{a,lstm\text{-}tl}$ outperformed $wtd_{a,lstm(m)}$ and $wtd_{a,m}$ at peaks ($|wtd_{a,o}| > 0.5$), obtaining increases of 0.12 and 0.14 in NSE, which reflects the added value of LSTM-TL in predicting extreme events such as GW droughts.

4.3.2 Recent groundwater drought analysis using reconstructed pan-European long-term water table depth anomaly data

The time length of $wtd_{a,lstm\text{-}tl}$ is determined by input $pr_{a,o}$ and $\theta_{a,o}$, which allows the generation of $wtd_{a,lstm\text{-}tl}$ beyond the time period of $wtd_{a,m}$ used for training (i.e., the 1996-2016 period). As aforementioned, this is an advantage of LSTM-TL. Here we adopted LSTM-TL to reconstruct pan-European monthly wtd_a data RD1-6 (Table 4.1) from the early 1980s to the near present based on various $pr_{a,o}$ and $\theta_{a,o}$ pairs from observational datasets (i.e., $pr_{a,o}$ from the ERA5 bias-corrected, COSMO-REA6 and ERA5-Land datasets and $\theta_{a,o}$ from the ERA5-Land and GLEAM datasets). Different input $pr_{a,o}$ and $\theta_{a,o}$ data result in different $wtd_{a,lstm\text{-}tl}$ data. The internal LSTM networks were only trained once on modeling results and then used for data reconstruction without additional training.

Table 4.1: Sources of input $pr_{a,o}$ and $\theta_{a,o}$ data utilized to reconstruct wtd_a data RD1-6 as well as time periods of RD1-6.

Reconstructions	Data source of input $pr_{a,o}$	Data source of input $\theta_{a,o}$	Time period
RD1	ERA5 bias-corrected	ERA5-Land	01/1981-12/2019
RD2	ERA5 bias-corrected	GLEAM	01/1980-12/2019
RD3	COSMO-REA6	ERA5-Land	01/1995-08/2019
RD4	COSMO-REA6	GLEAM	01/1995-08/2019
RD5	ERA5-Land	ERA5-Land	01/1981-05/2021
RD6	ERA5-Land	GLEAM	01/1981-12/2020

Figure 4.4 illustrates the spatial distribution of yearly averaged $wtd_{a,lstm\text{-}tl}$ over the European continent from RD1-6 for 2003, 2015, 2018 and 2019, constituting drought years (Fink et al., 2004; Van Lanen et al., 2016; Bastos et al., 2020; Boergens et al., 2020). The spatial distribution presents similar patterns of positive and negative $wtd_{a,lstm\text{-}tl}$ over Europe across all reconstructed data for the investigated years. We found positive anomalies (i.e., dry events) in all European regions except Iberian Peninsula (IP) for 2003 and in ME, EA, parts of France (FR) and IP for 2015, which agree well with the reported regions affected by meteorological droughts (Fink et al., 2004; Van Lanen et al., 2016). During the 2018-2019 period, the central and northeastern parts of Europe experienced persistent GW droughts. The GW deficits were triggered by extreme spring warming and brightening in May 2018 and further exacerbated during consecutive summer heat waves in 2018 and 2019 (Bastos et al., 2020; Boergens et al., 2020; Hari et al., 2020). A recovery of GW storage took more than one year in ME, EA and FR (Figure 4.5). In these regions, the strongest impact

of GW drought (manifested by the maximum $wtd_{a,lstm\text{-}tl}$) was observed in October and November 2018 after summer precipitation deficits, reflecting the lagged response of GW to precipitation.

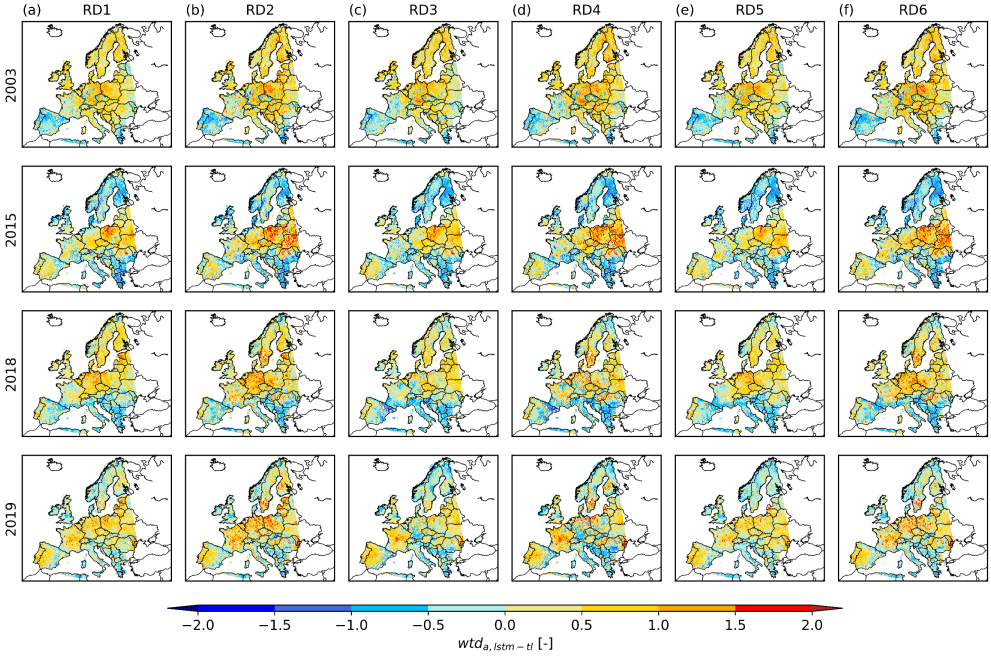

Figure 4.4: Spatial distribution of yearly averaged $wtd_{a,lstm\text{-}tl}$ over the European continent for the recent drought years 2003, 2015, 2018 and 2019. The anomalies were calculated from (a) RD1; (b) RD2; (c) RD3; (d) RD4; (e) RD5; and (f) RD6.

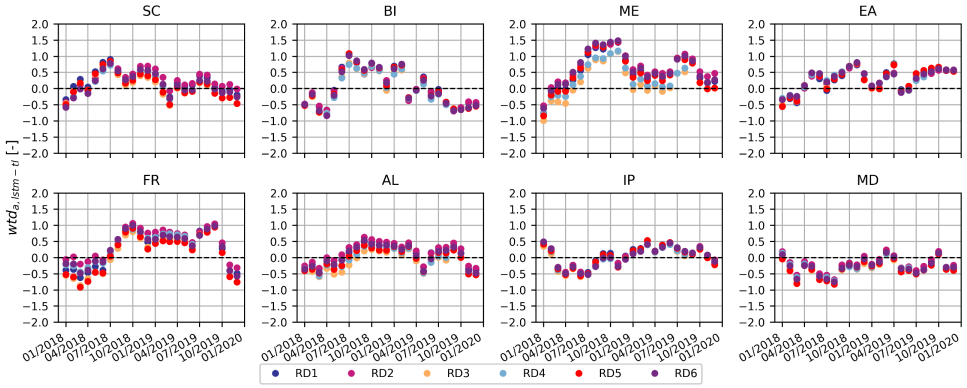

Figure 4.5: Regionally averaged $wtd_{a,lstm\text{-}tl}$ from RD1-6 over eight PRUDENCE regions for the 2018-2019 period. The black dashed lines indicate $wtd_a = 0$. A long time period with $wtd_a > 0$ corresponds to a dry event while a long time period with $wtd_a < 0$ corresponds to a wet event.

4.3.3 Seasonal trends of water table depth anomalies in various European regions

We derived seasonal wtd_a trends in different PRUDENCE regions by fitting seasonal averaged $wtd_{a,lstm\text{-}tl}$ from the reconstructed data RD1-6 into first-order polynomials y = mx + b. The m is the change rate of wtd_a [y^{-1}], where m > 0 corresponds to a drying tend and m < 0 corresponds to a wetter condition. The Wald test (Hauck and Donner, 1977) was utilized to determine the statistical significance of the trends, with 95% as the threshold. For verification, we compared seasonal averaged wtd_a and their trends in ME (i.e., the PRUDENCE regions with the most pixels with $wtd_{a,o}$) obtained from RD1-6 and $wtd_{a,o}$ for the 1996-2016 period (Figure 4.6). The seasonal averaged $wtd_{a,lstm\text{-}tl}$ from RD1-6 showed a good match with $wtd_{a,o}$ in ME except overestimated values in the Summer-Winter 2002 (a record flood over Europe). Both RD1-6 and $wtd_{a,o}$ revealed that there was no significant wtd_a trend in ME for the 1996-2016 period.

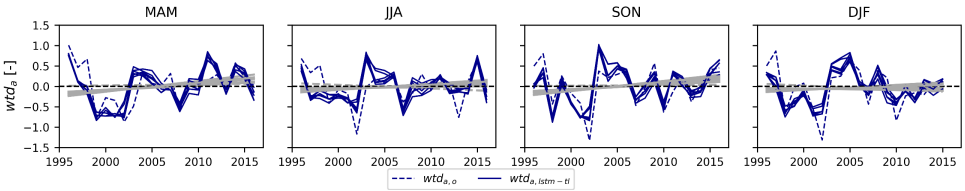

Figure 4.6: Seasonal averaged wtd_a and their trends in ME from RD1-6 and $wtd_{a,o}$ for the 1996-2016 period (MAM - March, April and May, i.e., spring; JJA - June, July and August, i.e., summer; SON - September, October and November, i.e., autumn; DJF - December, January and February, i.e., winter). The bold lines show the seasonal trends of $wtd_{a,lstm\text{-}tl}$ in RD1-6 while the bold dashed lines present the seasonal trends of $wtd_{a,o}$. The red, blue and gray trend lines indicate a drier GW condition at the 95% significant level, a wetter GW condition at the 95% significant level and no significant trend, respectively. The black dashed lines represent $wtd_a = 0$. A long time period with $wtd_a > 0$ corresponds to a dry event while a long time period with $wtd_a < 0$ corresponds to a wet event.

Figure 4.7 maps seasonal wtd_a trends at the European scale in RD1-6 for the time periods listed in Table 4.1, with pixels without significant trends masked in gray. Inspecting the maps, ME, EA and FR tended to experience intensified GW droughts while parts of MD tended towards a wetter GW pattern during the study periods. This finding is basically consistent with the trend pattern observed in the GRACE terrestrial water storage data for the 2002-2017 period, that is, a distinct pattern of mid-latitude drying and of high- and low-latitude wetting (Reager et al., 2016; Tapley et al., 2019). However, the statistical significance of the GRACE-observed trend pattern was not assessed. The seasonal wtd_a trends are strongly influenced by the LSTM-TL input $pr_{a,o}$ and $\theta_{a,o}$. The reconstructed data by LSTM-TL using GLEAM $\theta_{a,o}$ as input (RD2, RD4 and RD6, Figures 4.7b, d and f) exhibited more pronounced and coherent wtd_a trends in four seasons than the ones using ERA5-Land $\theta_{a,o}$ as input (RD1, RD3 and RD5, Figures 4.7a, c and e). The latter showed weaker wtd_a trends in spring and winter. In general, the seasonal wtd_a trends in the reconstructed data were in good agreement with the trends in the LSTM-TL input $pr_{a,o}$ and $\theta_{a,o}$ data used to produce the data (Figures C.3-8).

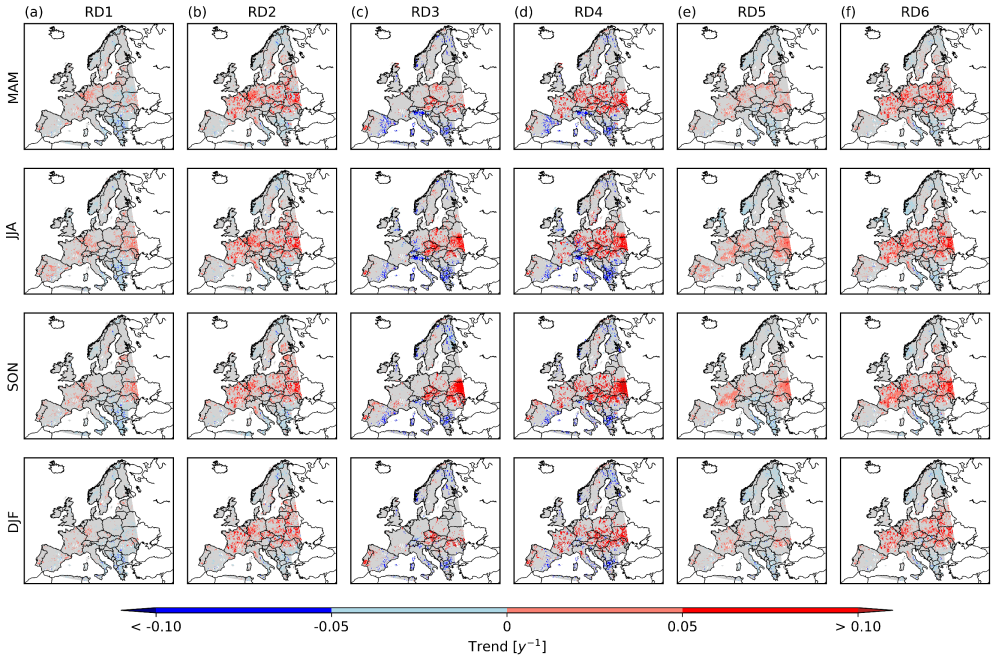

Figure 4.7: Seasonal *wtd*$_a$ trend maps over Europe (MAM - March, April and May, i.e., spring; JJA - June, July and August, i.e., summer; SON - September, October and November, i.e., autumn; DJF - December, January and February, i.e., winter). The seasonal *wtd*$_a$ trends were derived from (a) RD1; (b) RD2; (c) RD3; (d) RD4; (e) RD5; and (f) RD6. The pixels without significant trends are masked in gray.

4.4 Discussion

Efficient GW monitoring is essential for European GW management, especially under increasing pressure with changing climate. Yet, it is challenging to monitor GW at large scales, mainly due to the lack of *wtd* observations. To address this challenge, here we propose the LSTM-TL methodology, which combines LSTM networks and TL, to produce spatiotemporally continuous *wtd*$_a$ estimates at the European scale in the absence of *wtd* observations. The core idea of the methodology is to transfer the modeled relationship (f_m) between *wtd*$_a$ and input hydrometeorological variables (here *pr*$_a$ and θ_a) to the observation-based estimation, in order to provide reliable *wtd*$_a$ estimates for regions without *wtd* observations. The f_m is assumed to agree with the observed one (f_o). Compared with physically-based numerical models, LSTM-TL requires orders of magnitude less computational resources in generating *wtd*$_a$ estimates (i.e., *wtd*$_{a,lstm\text{-}tl}$) at the same temporal and spatial scales, thereby facilitating early warning and fast decision making for extreme GW events. Based on the European monthly *wtd*$_a$ data reconstructed by LSTM-TL, we provided the first estimates of seasonal *wtd*$_a$ trends in different European regions, which show that central and eastern Europe had significant drying trends during the study periods.

Some limitations of the derived LSTM-TL should be recognized. First, in this study, LSTM-TL neglects anthropogenic interventions, such as GW pumping, on GW dynamics, because the modeling results used for training the internal LSTM networks give a near-natural representation of the water cycle (Furusho-Percot et al., 2019). Thus, the methodology may underestimate wtd_a in regions with strong human impacts such as IP and MD (de Graaf et al., 2019). Second, while the utilized modeling results guarantee good spatial coverage of the European domain, their data are limited at the individual pixel level (< 300 time steps for training), which may impair the skill of LSTM-TL in estimating wtd_a. Improved performance of LSTM-TL can be expected as training data increase. Last but not least, we found the LSTM networks in LSTM-TL usually had poor test performance in deep aquifers, which might be attributed to the inability to capture very long-term dependences between wtd_a and input hydrometeorological variables. Some alternative DL techniques that are superior in time series forecasting can be considered, such as Transformers (Vaswani et al., 2017), which can effectively capture dependences within the time steps used for training.

Despite the above limitations, our results show that LSTM-TL had reasonable performance over Europe and may serve as a useful alternative to *in-situ* $wtd_{a,o}$ until collated *wtd* data sets are available. In addition to data reconstruction described in this study, LSTM-TL can be deployed in online GW monitoring, which would be useful in European GW management.

Chapter 5 Conclusions and outlook

5.1 Summary and conclusions

Efficient groundwater (GW) monitoring is indispensable for GW management under today's changing climate and increasing human interventions. Up to date, however, it is still challenging to monitor GW at large scales, mainly due to the lack of collated water table depth (*wtd*) observations. The understanding of extreme GW events is often derived from meteorological data (e.g., precipitation), which is insufficient. Although alternative methodologies, such as physically-based numerical models, have been used to predict and quantify GW changes (see Section 1.3 for details), they are computationally expensive for high-resolution long-term estimation over large regions. Therefore, this PhD work proposes a novel methodology based on advanced machine learning (ML) techniques, i.e., Long Short-Term Memory (LSTM) networks and transfer learning (TL), to produce reliable monthly *wtd* anomaly (wtd_a) estimates at the European scale with low computational cost, which is named LSTM-TL. The core idea of LSTM-TL is to generate wtd_a estimates from closely related hydrometeorological variables that have spatiotemporally continuous coverage over Europe. In this work, precipitation and soil moisture anomalies (pr_a and θ_a) are selected as input. Because of limited wtd_a observations ($wtd_{a,o}$) available for training, LSTM-TL incorporates established knowledge from an uncalibrated physically-based modeling system (i.e., Terrestrial Systems Modeling Platform, TSMP) into the internal LSTM networks, assuming that the modeled input-output relationship (f_m) agrees well with the observed (f_o). The construction of the internal LSTM networks is merely based on the TSMP-G2A data set, which encompasses simulation results from TSMP. The steps of developing LSTM-TL are described in three chapters (Chapters 2 to 4), including exploring the potential of LSTM networks to estimate wtd_a over Europe, optimizing the developed LSTM networks and applying the optimized LSTM networks to reality. In the first step, the basic architecture of the LSTM networks in LSTM-TL and the generic workflow to train the LSTM networks at individual pixels are established. In the second step, the input variables of LSTM-TL are determined. In the final step, LSTM-TL is proposed for real-world applications. All the data utilized in this work are from the TSMP-G2A data set, except for the final step involving observations. The findings in the three steps are summarized as follows.

In the first step (presented in Chapter 2), the potential of LSTM networks to estimate wtd_a at the individual pixel level over Europe based on pr_a was investigated assuming that pr_a is a useful proxy for wtd_a. LSTM networks were applied because they are superior in exploiting the long-short-term dependencies between time sequences, which is expected in the lagged response of wtd_a to pr_a. The reproduced wtd_a maps by the LSTM results showed good agreement with the TSMP-G2A wtd_a maps for drought months in terms of the spatial distribution of dry and wet events, demonstrating the ability of LSTM networks to produce reliable wtd_a estimates with respect to the TSMP-G2A data set. The network test performance was significantly affected by local factors, including yearly averaged *wtd*, evapotranspiration (*ET*), soil moisture (θ), and snow water equivalent (*SWE*). This also resulted in varying network test performance in hydrometeorologically different regions over Europe. Cross-wavelet transform (XWT) was conducted to analyze the time-frequency patterns in the TSMP-G2A pr_a and wtd_a data, and the results confirmed that the inability of the LSTM

networks to estimate wtd_a at some pixels was due to changing temporal patterns in the TSMP-G2A pr_a-wtd_a relationship during the study period.

In the second step (presented in Chapter 3), the performance of the developed LSTM networks in the previous step was improved via introducing additional input variables. A number of input hydrometeorological variables were studied in various experiments, which are pr_a, *ET* anomaly (ET_a), θ_a, scaled yearly averaged *SWE* (SWE_{scaled}), and anomalies at adjacent pixels (e.g., river stage anomaly, rs_a). The LSTM networks with pr_a and θ_a as input achieved the most improvement in the wtd_a estimation at the European scale. Considering θ_a substantially improved the obtained wtd_a estimates especially in the areas with *wtd* ≤ 3 m (i.e., the major *wtd* category of Europe), and thus, it is important to integrate ground-based information such as θ with meteorological information to predict and quantify GW changes over Europe. As θ_a and wtd_a are measures of agricultural drought and GW drought, respectively, the considerable contribution of θ_a to the wtd_a estimation also indicates a close connection between agricultural drought and GW drought on the European continent.

In the final step (presented in Chapter 4), TL was utilized to transfer f_m between wtd_a and input hydrometeorological variables (here pr_a and θ_a) to the observation-based estimation of wtd_a, and LSTM-TL was proposed to provide reliable wtd_a estimates for regions with limited or without *wtd* observations. The methodology constitutes of two steps, that are training the internal LSTM networks on the data derived from the TSMP-G2A data set and estimating wtd_a based on the observed pr_a and θ_a ($pr_{a,o}$ and $\theta_{a,o}$) and the relationship learned from the TSMP-G2A data set by the internal LSTM networks (f_{lstm}, assuming $f_{lstm} \approx f_m$). LSTM-TL obtained wtd_a estimates in good agreement with *in-situ* $wtd_{a,o}$ at approximately 2,600 European GW monitoring wells, showing the reliability of the methodology. LSTM-TL can produce wtd_a estimates beyond the time period of the TSMP-G2A data set used for training, where the time period of the obtained wtd_a estimates is determined by input $pr_{a,o}$ and $\theta_{a,o}$. Using LSTM-TL, monthly pan-European wtd_a data from the early 1980s to the near present were reconstructed based on various $pr_{a,o}$ and $\theta_{a,o}$ pairs. The reconstructed wtd_a data provided the first estimate of seasonal wtd_a trends in different European regions in the past, contributing significantly to the understanding of historical GW dynamics at the continental scale over Europe. Compared with physically-based numerical models, LSTM-TL requires orders of magnitude less computational resources in estimating wtd_a at the same temporal and spatial scales, thus advancing early warning and fast decision making for extreme GW events.

The results of this PhD work demonstrate the ability of LSTM networks and LSTM-TL to generate reliable wtd_a estimates at the European scale with appropriate input variables. During the development of LSTM-TL, the structure of the internal LSTM networks, the processes of constructing the internal LSTM networks, the data used to train the internal LSTM networks (here the TSMP-G2A data set) and the input variables of LSTM-TL have been determined. Thus, the uncertainties in the obtained wtd_a estimates by LSTM-TL mainly come from input $pr_{a,o}$ and $\theta_{a,o}$ data. While LSTM-TL has been implemented over Europe, it can be transferred to other regions in the world by changing the simulation data used to build the internal LSTM networks and

adjusting its input variables. This work does not conclude that the role of physically-based numerical models in the hydrometeorological sciences can be replaced with ML techniques, which provide important data in the absence of ground-based observations, e.g., *wtd*. The LSTM networks in LSTM-TL are trained on simulation results from physically-based numerical models at individual pixels over Europe, and the success of LSTM-TL in estimating wtd_a highlights the advantage of combining ML techniques with knowledge contained in physically-based numerical models in hydrological studies.

5.2 Limitations

ML-based models are constructed exclusively from data, and thus, their performance is strongly influenced by the quality and quantity of the data utilized to build the model. Compared to shallow artificial neural networks (ANNs), deep learning (DL) techniques including LSTM networks generally require more data for training, due to their more complex and larger architectures. While in this study, the TSMP-G2A data set guarantees spatiotemporally continuous coverage over Europe, it provides limited data (< 300 time steps) at the individual pixel level for constructing the internal LSTM networks, which may constrain the performance of LSTM-TL. The TSMP-G2A data set represents a near-natural climatology of the physical states of the terrestrial water and energy cycles, so the impact of human interventions is neglected in the current LSTM-TL implementation. The obtained wtd_a estimates may be biased in regions with intensive human activities such as reservoirs and farmlands with GW abstractions and irrigation. As a result of anthropogenic warming, recent decades have seen many record-breaking extreme events worldwide (Guerreiro et al., 2018), e.g., the European summer heat wave of 2003 (Fink et al., 2004) and the consecutive European summer heat waves and droughts of 2018 and 2019 (Boergens et al., 2020). Hence, historical observations may not always be a reliable guide for future events, further limiting the predictive capacities of LSTM-TL driven completely by historic data.

Hyperparameter tuning is an important step in the construction of ML-based models. Hyperparameters refer to parameters that govern the training process of ML-based models, such as the optimizer and its learning rate, initial states of weights and biases, and number of hidden neurons in an ANN. The ranges of hyperparameters to be tuned are problem-specific and depend on user's experience. Tuning a large number of hyperparameters is usually computational demanding. Here, to save computational resources, only the number of hidden neurons was tuned, ranging from 1 to 100, which may reduce the ability of LSTM-TL to estimate wtd_a.

The interpretability of a methodology stands for the ability to explain its performance, which is critical in hydrological sciences. The use of ML-based models is often criticized for their "black-box-ness". In particular, the behavior of LSTM networks is more difficult to interpret than other ML techniques because of the time-varying weights and biases in the linkages between neurons. This is also the case in the presented study.

ML-based models only address unidirectional cause-effect chains. The LSTM networks utilized in this work consider the impact of input hydrometeorological forcings on target wtd_a, but neglect the reverse influence (detailed in Section 1.2). The resultant disagreement between the estimated and observed wtd_a requires further investigation.

Furthermore, the LSTM networks in LSTM-TL tended to perform poorly in deep aquifers, resulting in less reliable wtd_a estimates. The poor performance may be attributed to the small training data size and simplified hyperparameter tuning, and it may also suggest the inability of LSTM networks to exploit very long-term dependencies between input and target time sequences.

While considering the above limitations, LSTM-TL has been shown to generate reliable wtd_a estimates over Europe and may serve as an additional methodology to *in-situ* $wtd_{a,o}$ and time series analyses..

5.3 Outlook

LSTM-TL is useful in practical applications, such as online GW monitoring and GW drought predictions based on e.g. medium-range and seasonal weather and soil moisture forecasts. Because of sparse *wtd* observations over large regions, the obtained wtd_a estimates can be utilized for validation of *wtd* products. Although the methodology has been proposed for estimating wtd_a over Europe, it can be adapted to other regions and ground-based variables that lack large-scale observations. As illustrated by the XWT analysis, the poor performance of the LSTM networks at some pixels was induced by changing temporal patterns in the TSMP-G2A pr_a-wtd_a relationship during the study period. The changing temporal patterns may indicate signals of climate change and should be further analyzed.

In the future, LSTM-TL can be improved in terms of predictive performance and interpretability. Increasing training data size and fine-tuning hyperparameters are expected to improve the obtained wtd_a estimates. During hyperparameter tuning, a trade-off should be made between reliable estimates and reasonable computational cost. Remotely sensed observations of ground-based variables, such as GRACE terrestrial water storage anomalies (Watkins et al., 2015), can be included in LSTM-TL as input to account for human impacts on GW dynamics. Additionally, the architecture of the internal LSTM networks can be modified to address the two-way feedbacks between wtd_a and input hydrometeorological variables, which may lead to improved wtd_a estimates. Moreover, Temporal Convolutional Networks (Yan et al., 2020) and Transformers (Vaswani et al., 2017) can be substitutes for the internal LSTM networks, which have been shown to outperform LSTM networks in long-term time series prediction.

With the advances in ML/DL, many ML/DL interpretation methods are available for explaining LSTM networks' results, e.g., integrated gradients (Sundararajan et al., 2017), contextual decomposition (Murdoch et al., 2018) and Layer-wise Relevance Propagation (Arras et al., 2019). A good understanding of the LSTM-TL behavior will further improve its output. In addition, some more transparent ML/DL techniques can be applied. Transformers are also a good option in this context.

5.4 Code and data availability

The python scripts for constructing LSTM networks and LSTM-TL and analyzing results are available online at https://icg4geo.icg.kfa-juelich.de/SoftwareTools/GEOEssential/-/tree/master/Scripts_papers. The TSMP-G2A data set is available online at https://doi.org/10.17616/R31NJMH3 (Furusho-Percot et al., 2019). The availability of *in-situ wtd* measurements utilized in the evaluation of LSTM-TL is stated in Table C.1. The input, output and target data of LSTM networks and LSTM-TL related to this work are available online at the links listed in Table 5.1.

Table 5.1: Links of the input, output and target data of LSTM networks and LSTM-TL related to the PhD work.

Data	Link	Used in Chapter(s)
TSMP-G2A pr_a and wtd_a	https://doi.org/10.26165/JUELICH-DATA/WPRA1F	Chapters 2, 3 and 4
TSMP-G2A ET_a and SWE	https://doi.org/10.26165/JUELICH-DATA/AMQ6NI	Chapters 3
TSMP-G2A θ_a	https://doi.org/10.26165/JUELICH-DATA/AMQ6NI	Chapters 3 and 4
The wtd_a estimates obtained by LSTM networks of E1.1 (see Table 3.2)	https://doi.org/10.26165/JUELICH-DATA/WPRA1F	Chapters 2 and 3
The wtd_a estimates obtained by LSTM networks of E1.2 to E2.3 (see Table 3.2)	https://doi.org/10.26165/JUELICH-DATA/AMQ6NI	Chapter 3
Data related to the evaluation of the LSTM-TL performance*	https://doi.org/10.26165/JUELICH-DATA/ZBLDIR	Chapter 4
Reconstructed European monthly wtd_a data RD1-6 and their input $pr_{a,o}$ and $\theta_{a,o}$ data	https://doi.org/10.26165/JUELICH-DATA/ZBLDIR	Chapter 4

* Including input $pr_{a,o}$ and $\theta_{a,o}$, $wtd_{a,lstm\text{-}tl}$ (wtd_a estimates from LSTM-TL), $wtd_{a,lstm(m)}$ (wtd_a estimates obtained by LSTM networks using modeling results as input) and $wtd_{a,lstm(o)}$ (wtd_a estimates obtained by LSTM networks trained on observations) for the period 1996 to 2016.

Appendix

Appendix A Supplementary material to Chapter 2

Pseudocode of the LSTM network (displayed in Figure 2.2)

The pseudocode, shown below, is of the one-hidden-layer LSTM networks illustrated in Figure 2.2, which is modified from Gers et al. (2000). Variables were defined in the caption of Figure 2.2. Note that, in order to simplify the code, the biases are not shown here.

RESET all network parameters (i.e., weights, biases and cell states) as listed in Table 2.2
REPEAT learning loop
  forward pass
    for t = 1, 2, …
      network input to the hidden layer (self-recurrent and from input):
        input gate: $net_{in}(t) = w_{in}x(t) + w_{inh}h(t-1)$
        forget gate: $net_{forget}(t) = w_{forget}x(t) + w_{forgeth}h(t-1)$
        output gate: $net_{out}(t) = w_{out}x(t) + w_{outh}h(t-1)$
        cell: $net_c(t) = w_c x(t) + w_{ch}h(t-1)$
      activations in the hidden layer:
        input gate: $i(t) = \sigma(net_{in}(t))$
        forget gate: $f(t) = \sigma(net_{forget}(t))$
        output gate: $o(t) = \sigma(net_{out}(t))$
        cell's internal state:
        $c(0) = 0, c(t) = f(t)c(t-1) + i(t)g(t)$, where $g(t) = \tanh(net_c(t))$
        cell' s activation: $h(t) = o(t)\tanh(c(t))$
      output of the network:
        $net(t) = w_{net}h(t)$, out(t) = net(t)
  backward pass if error injected
    for t = n, n-1, …
      use RMSprop optimization algorithm (Hinton et al., 2012)
UNTIL validation error begins to drop and number of epochs ≥ 50

Additional European water table depth anomaly maps

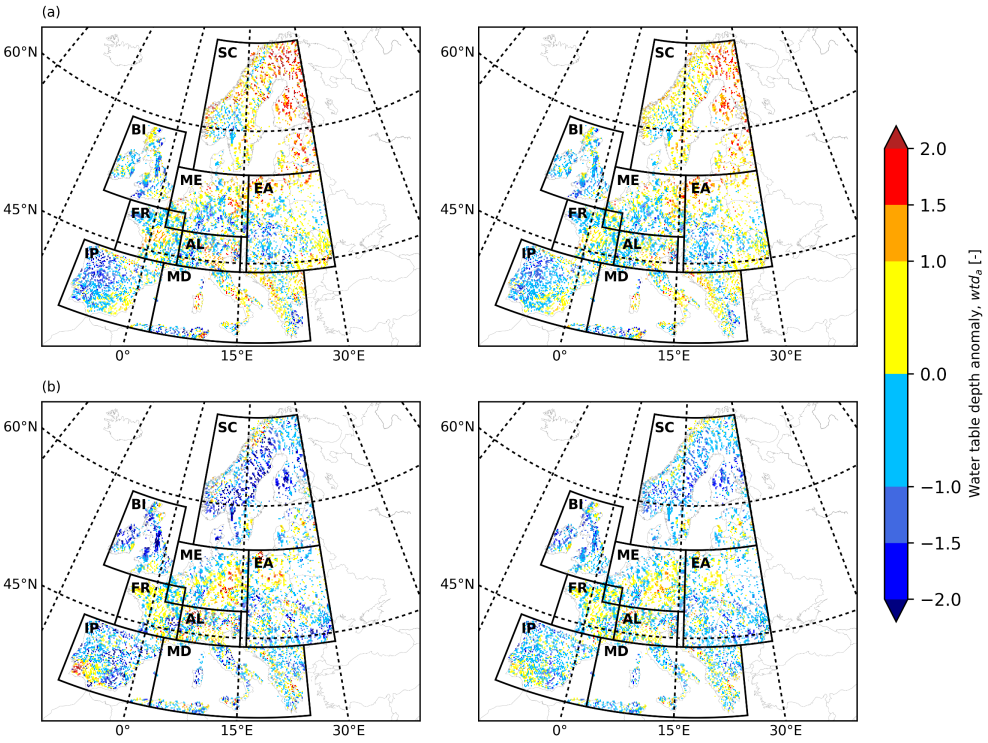

Figure A.1: European wtd_a maps for (a) July 2003 (i.e., in the training period) and (b) July 2015 (i.e., in the testing period) derived from the TSMP-G2A data set (left) and the results from LSTM networks (right).

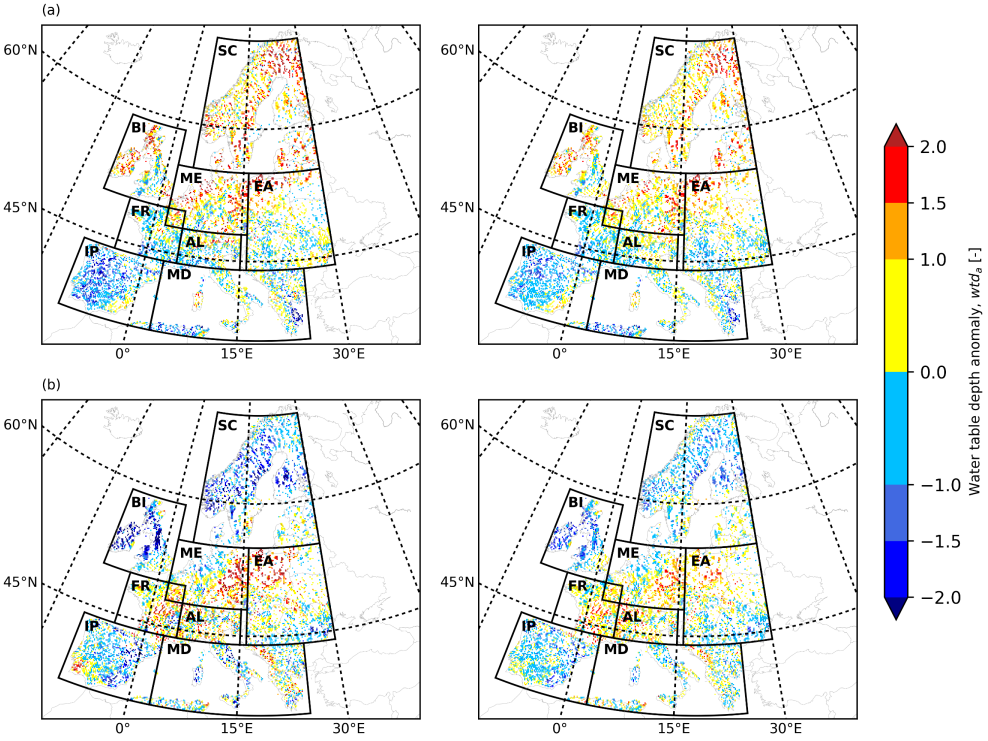

Figure A.2: European wtd_a maps for (a) December 2003 (i.e., in the training period) and (b) December 2015 (i.e., in the testing period) derived from the TSMP-G2A data set (left) and the results from LSTM networks (right).

Results of the cross-wavelet transform (XWT) analysis at additional pixels

Table A.1: Pixel characteristics in the XWT analysis (pixels 3-4).

	Performance combination	Region	Yearly averaged water table depth, *wtd* [m]	Yearly averaged evapotranspiration, *ET* [mm]	Yearly averaged soil moisture, θ [m^3 m^{-3}]	Yearly average snow water equivalent, *SWE* [mm]
Pixel 3	C1	FR	1.06	418.39	0.31	0.0
Pixel 4	C2	IP	6.44	153.92	0.16	0.0
	Training NSE [%]	Training α [%]	Validation NSE [%]	Validation α [%]	Test NSE [%]	Test α [%]
Pixel 3	84.29	97.89	60.61	98.38	62.22	84.87
Pixel 4	94.39	99.79	46.87	90.86	-724.90	20.26

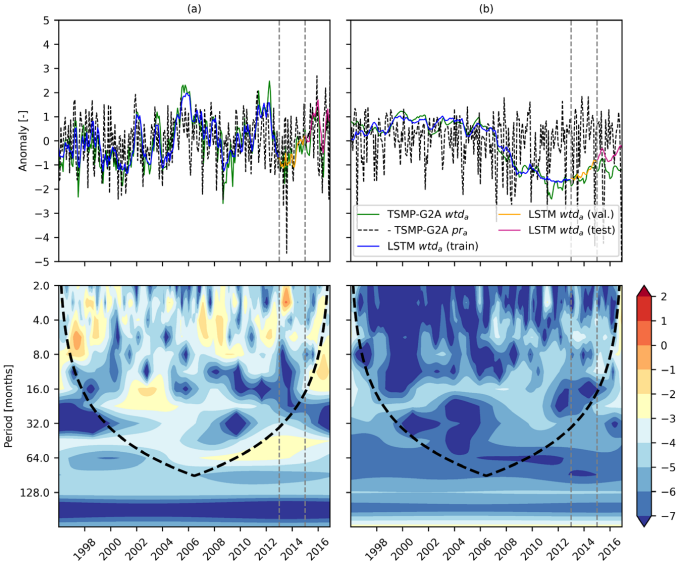

Figure A.3: TSMP-G2A pr_a, TSMP-G2A wtd_a, and LSTM wtd_a time series (top), as well as cross-wavelet spectra for TSMP-G2A pr_a and wtd_a series (bottom), at (a) pixel 3 and (b) pixel 4. The lines here have the same definitions as in Figure 2.9.

Appendix B Supplementary material to Chapter 3

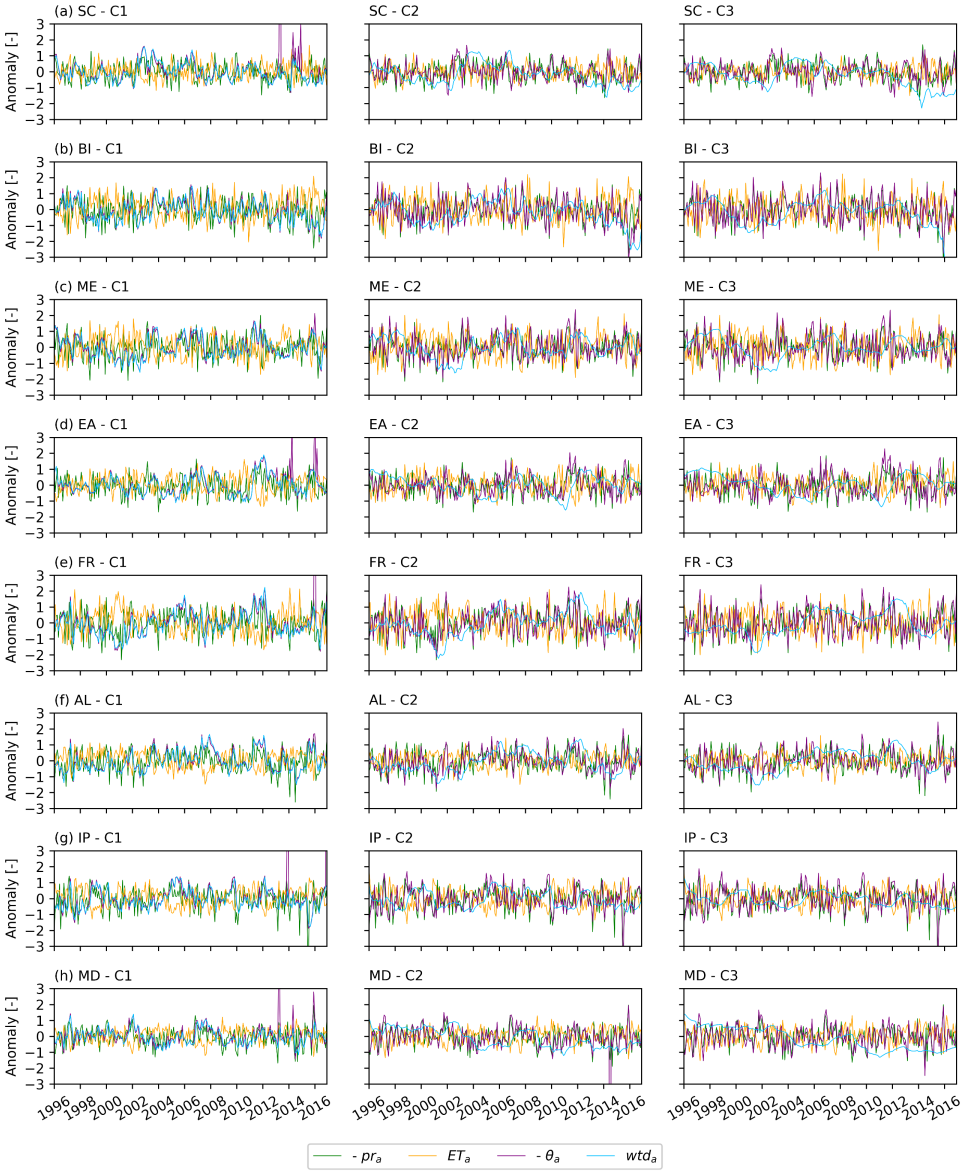

Figure B.1: Regionally averaged pr_a, ET_a, θ_a and wtd_a time series for the *wtd* categories C1 to C3 in different PRUDENCE regions.

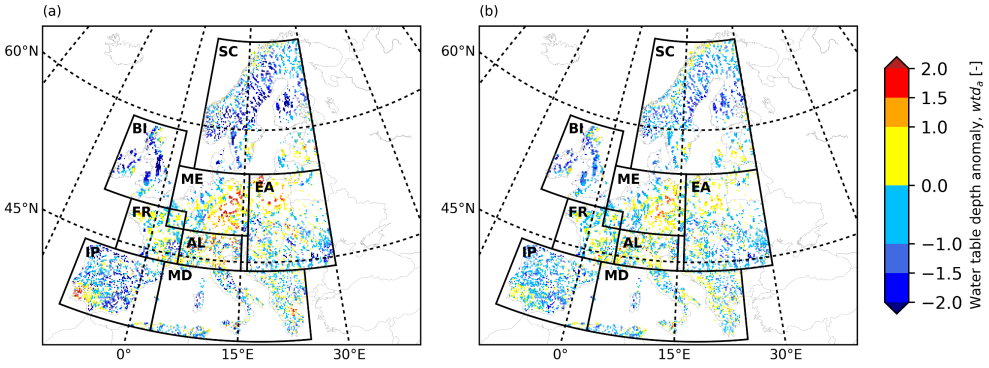

Figure B.2: European wtd_a maps for August 2015 (i.e., in the testing period) derived from (a) the TSMP-G2A data set and (b) the results from the LSTM networks of E1.1 (pr_a).

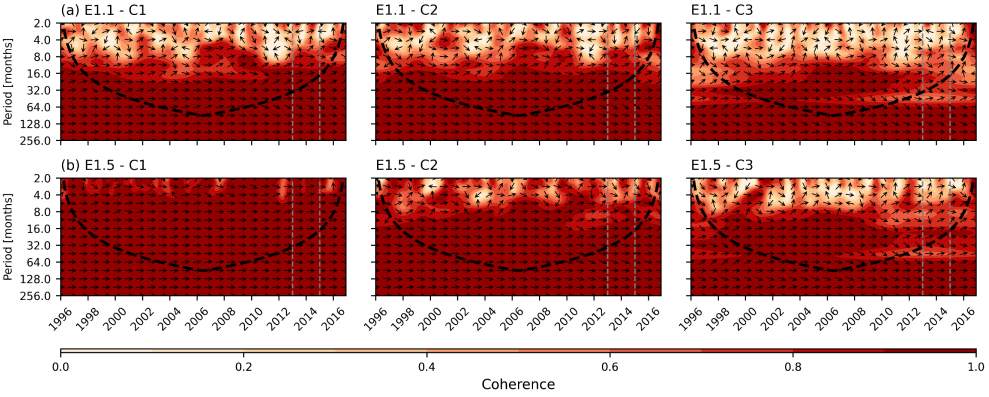

Figure B.3: Results of wavelet coherence analysis on the regionally averaged wtd_a time series for the *wtd* categories C1 to C3 in SC, which were derived from the TSMP-G2A data set and the results of the LSTM networks of: (a) E1.1: pr_a; and (b) E1.5: pr_a and θ_a.

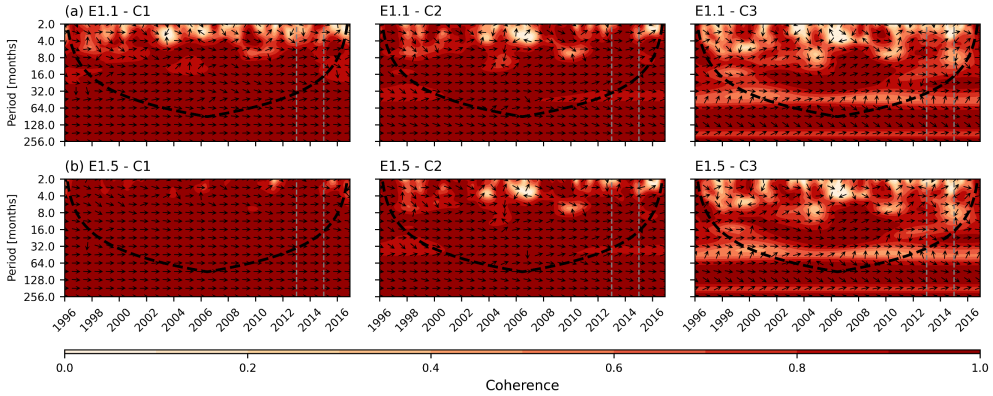

Figure B.4: Results of wavelet coherence analysis on the regionally averaged wtd_a time series for the *wtd* categories C1 to C3 in BI, which were derived from the TSMP-G2A data set and the results of the LSTM networks of: (a) E1.1: pr_a; and (b) E1.5: pr_a and θ_a.

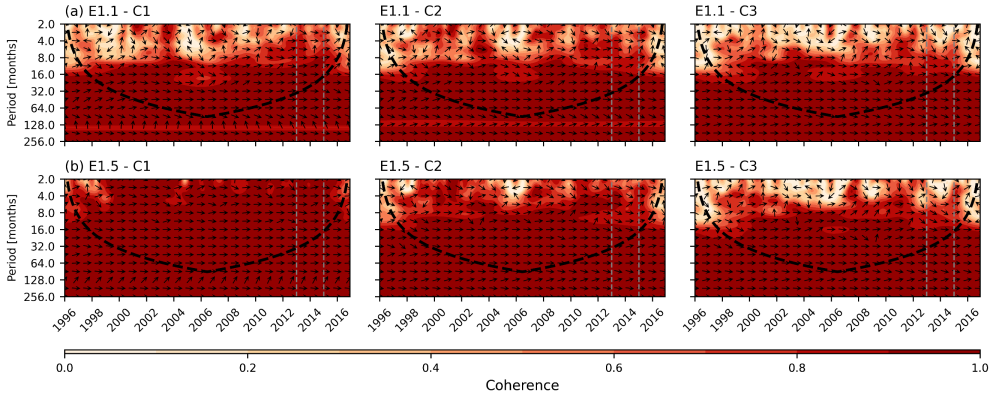

Figure B.5: Results of wavelet coherence analysis on the regionally averaged wtd_a time series for the *wtd* categories C1 to C3 in EA, which were derived from the TSMP-G2A data set and the results of the LSTM networks of: (a) E1.1: pr_a; and (b) E1.5: pr_a and θ_a.

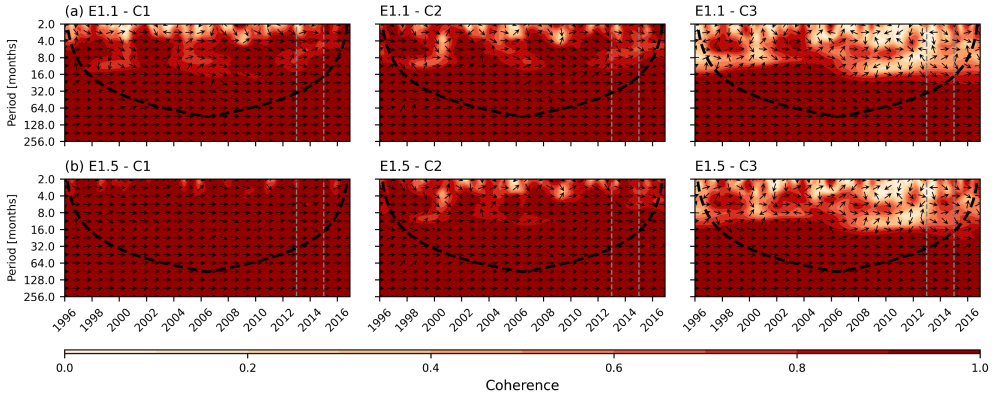

Figure B.6: Results of wavelet coherence analysis on the regionally averaged wtd_a time series for the *wtd* categories C1 to C3 in FR, which were derived from the TSMP-G2A data set and the results of the LSTM networks of: (a) E1.1: pr_a; and (b) E1.5: pr_a and θ_a.

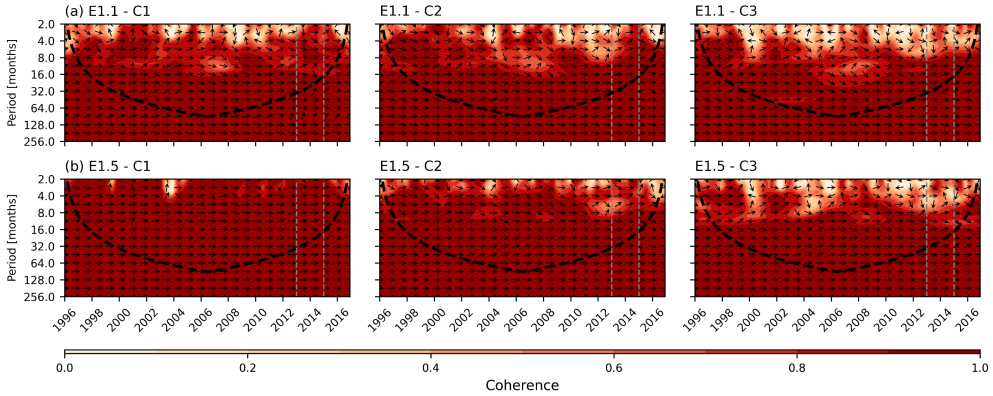

Figure B.7: Results of wavelet coherence analysis on the regionally averaged wtd_a time series for the *wtd* categories C1 to C3 in AL, which were derived from the TSMP-G2A data set and the results of the LSTM networks of: (a) E1.1: pr_a; and (b) E1.5: pr_a and θ_a.

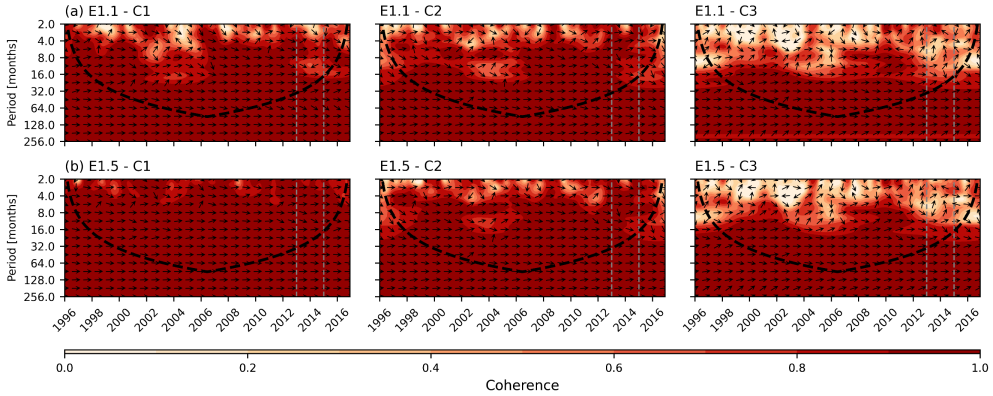

Figure B.8: Results of wavelet coherence analysis on the regionally averaged wtd_a time series for the *wtd* categories C1 to C3 in IP, which were derived from the TSMP-G2A data set and the results of the LSTM networks of: (a) E1.1: pr_a; and (b) E1.5: pr_a and θ_a.

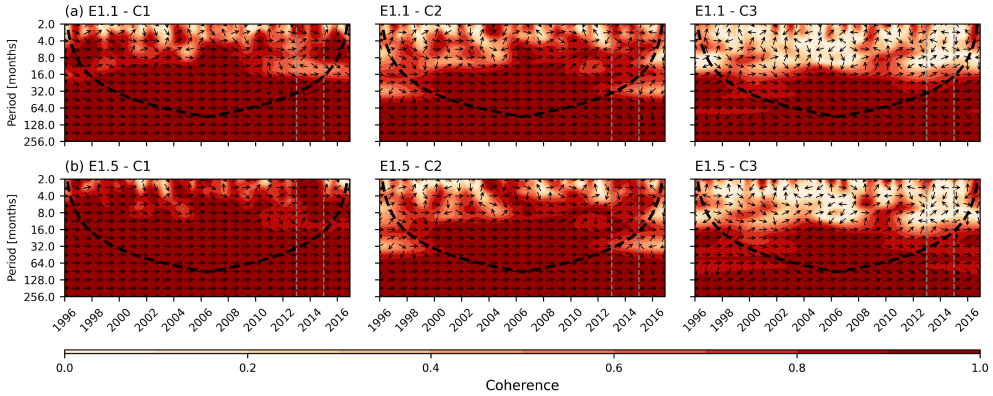

Figure B.9: Results of wavelet coherence analysis on the regionally averaged wtd_a time series for the *wtd* categories C1 to C3 in MD, which were derived from the TSMP-G2A data set and the results of the LSTM networks of: (a) E1.1: pr_a; and (b) E1.5: pr_a and θ_a.

Table B.1: Medians of the test NSE scores and RMSEs achieved in different experiments for C1 to C3 in each PRUDENCE region (set negative NSE as zeros).

Region	Experiment	Test NSE [%]			Test RMSE [-]			Region	Experiment	Test NSE [%]			Test RMSE [-]		
		C1	C2	C3	C1	C2	C3			C1	C2	C3	C1	C2	C3
SC	E1.1	13.30	0.00	0.00	0.57	0.6	0.87	FR	E1.1	74.96	44.82	20.52	0.40	0.34	0.37
	E1.2	0.00	0.00	0.00	0.84	1.12	1.37		E1.2	20.34	0.00	0.00	0.70	0.59	0.71
	E1.3	74.52	0.00	0.00	0.29	0.76	1.005		E1.3	94.41	42.06	0.00	0.19	0.37	0.43
	E1.4	13.39	0.00	0.00	0.58	0.64	0.885		E1.4	73.72	36.35	0.00	0.39	0.37	0.41
	E1.5	79.25	0.00	0.00	0.28	0.58	0.78		E1.5	95.03	57.44	13.97	0.17	0.32	0.36
	E1.6	76.77	0.00	0.00	0.29	0.76	0.965		E1.6	92.77	41.97	0.00	0.21	0.38	0.47
	E1.7	78.55	0.00	0.00	0.27	0.6	0.79		E1.7	93.50	51.47	8.91	0.19	0.33	0.38
	E2.1	79.14	0.00	0.00	0.27	0.59	0.8		E2.1	94.66	58.15	9.79	0.17	0.32	0.40
	E2.2	75.58	15.19	0.00	0.29	0.53	0.735		E2.2	92.40	65.88	15.52	0.20	0.27	0.37
	E2.3	79.45	12.07	0.00	0.28	0.55	0.74		E2.3	93.84	65.36	9.51	0.20	0.31	0.42
BI	E1.1	49.98	12.6	0.00	0.51	0.69	1.00	AL	E1.1	51.38	40.09	0.00	0.52	0.45	0.41
	E1.2	0.00	0.0	0.00	0.77	1.46	1.56		E1.2	13.61	0.00	0.00	0.74	0.76	0.62
	E1.3	87.93	18.9	0.00	0.24	0.63	0.98		E1.3	87.52	46.65	0.00	0.28	0.44	0.46
	E1.4	47.91	0.0	0.00	0.53	0.77	1.06		E1.4	56.72	28.40	0.00	0.50	0.51	0.43
	E1.5	88.41	32.1	0.00	0.24	0.60	0.94		E1.5	89.42	54.58	0.00	0.25	0.38	0.41
	E1.6	86.98	20.5	0.00	0.26	0.66	0.98		E1.6	87.52	38.94	0.00	0.27	0.47	0.46
	E1.7	86.95	21.0	0.00	0.25	0.65	1.03		E1.7	89.50	47.94	0.00	0.25	0.40	0.41
	E2.1	88.70	28.8	0.00	0.23	0.64	0.94		E2.1	89.53	56.99	0.00	0.25	0.37	0.40
	E2.2	82.32	8.9	0.00	0.28	0.68	1.09		E2.2	87.17	53.00	0.00	0.28	0.43	0.43
	E2.3	86.94	29.2	0.00	0.25	0.60	0.64		E2.3	90.15	63.30	0.26	0.27	0.38	0.40
ME	E1.1	62.39	49.29	0.00	0.48	0.41	0.41	IP	E1.1	51.57	0.00	0.00	0.50	0.51	0.59
	E1.2	0.00	0.00	0.00	0.91	0.89	0.89		E1.2	13.62	0.00	0.00	0.62	0.62	0.68
	E1.3	90.21	38.57	0.00	0.25	0.45	0.45		E1.3	86.68	0.00	0.00	0.22	0.45	0.56
	E1.4	57.16	36.14	0.00	0.51	0.46	0.45		E1.4	56.95	0.00	0.00	0.44	0.49	0.59
	E1.5	90.94	58.15	0.00	0.22	0.38	0.38		E1.5	86.46	0.00	0.00	0.23	0.45	0.59
	E1.6	87.91	42.18	0.00	0.27	0.48	0.52		E1.6	84.93	0.00	0.00	0.24	0.48	0.64
	E1.7	89.27	54.26	0.00	0.25	0.39	0.41		E1.7	84.88	0.00	0.00	0.25	0.45	0.65
	E2.1	90.62	58.79	0.00	0.23	0.39	0.39		E2.1	86.56	0.00	0.00	0.22	0.45	0.58
	E2.2	87.23	62.96	12.03	0.27	0.34	0.38		E2.2	79.31	0.00	0.00	0.29	0.41	0.61
	E2.3	87.06	57.91	5.22	0.28	0.37	0.42		E2.3	83.62	7.88	0.00	0.26	0.42	0.48
EA	E1.1	41.53	11.18	0.00	0.47	0.45	0.47	MD	E1.1	43.01	0.00	0.00	0.49	0.42	0.37
	E1.2	0.00	0.00	0.00	0.64	0.78	0.74		E1.2	22.23	0.00	0.00	0.57	0.51	0.46
	E1.3	89.97	12.74	0.00	0.18	0.45	0.50		E1.3	70.98	0.00	0.00	0.32	0.43	0.38
	E1.4	45.51	0.00	0.00	0.46	0.48	0.49		E1.4	51.54	0.00	0.00	0.44	0.44	0.36
	E1.5	91.25	20.64	0.00	0.17	0.42	0.46		E1.5	75.97	0.00	0.00	0.30	0.40	0.37
	E1.6	88.44	12.92	0.00	0.20	0.46	0.55		E1.6	71.06	0.00	0.00	0.31	0.43	0.40
	E1.7	89.71	13.79	0.00	0.18	0.43	0.50		E1.7	74.53	0.00	0.00	0.30	0.42	0.39
	E2.1	90.70	26.44	0.00	0. 18	0.42	0.50		E2.1	75.56	0.00	0.00	0.30	0.41	0.39
	E2.2	86.86	31.88	0.00	0.21	0.37	0.50		E2.2	72.01	0.00	0.00	0.33	0.42	0.43
	E2.3	91.52	41.70	9.33	0.18	0.38	0.35		E2.3	76.97	0.47	0.00	0.29	0.38	0.31

Appendix C Supplementary material to Chapter 4

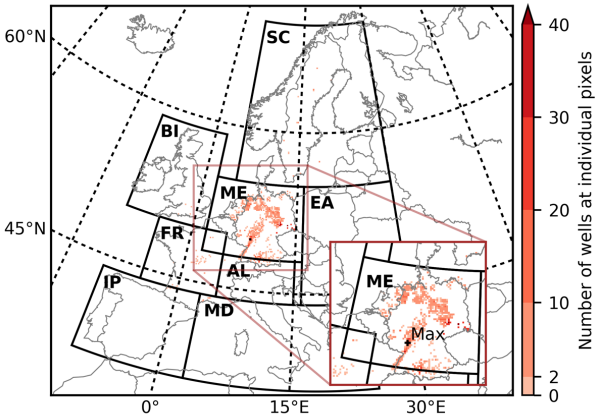

Figure C.1: Total number of GW monitoring wells located at individual pixels over Europe. The plus sign marks the pixel with the most GW monitoring wells (49).

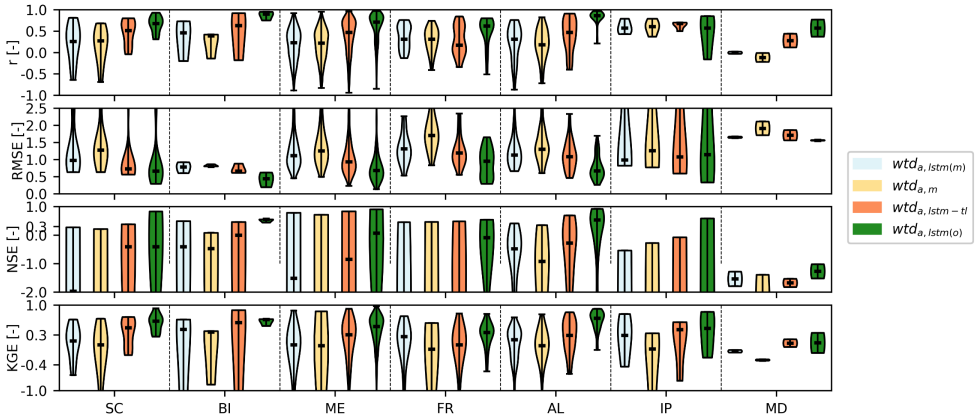

Figure C.2: Violin plots show r, RMSE, NSE, KGE of $wtd_{a,lstm(m)}$, $wtd_{a,m}$, $wtd_{a,lstm\text{-}tl}$ and $wtd_{a,lstm(o)}$ in seven PRUDENCE regions for the 2015-2016 period (i.e., the test period). The violin plots reveal the distribution of scores in each region, and the bold line in each violin plot indicates the regional median of scores.

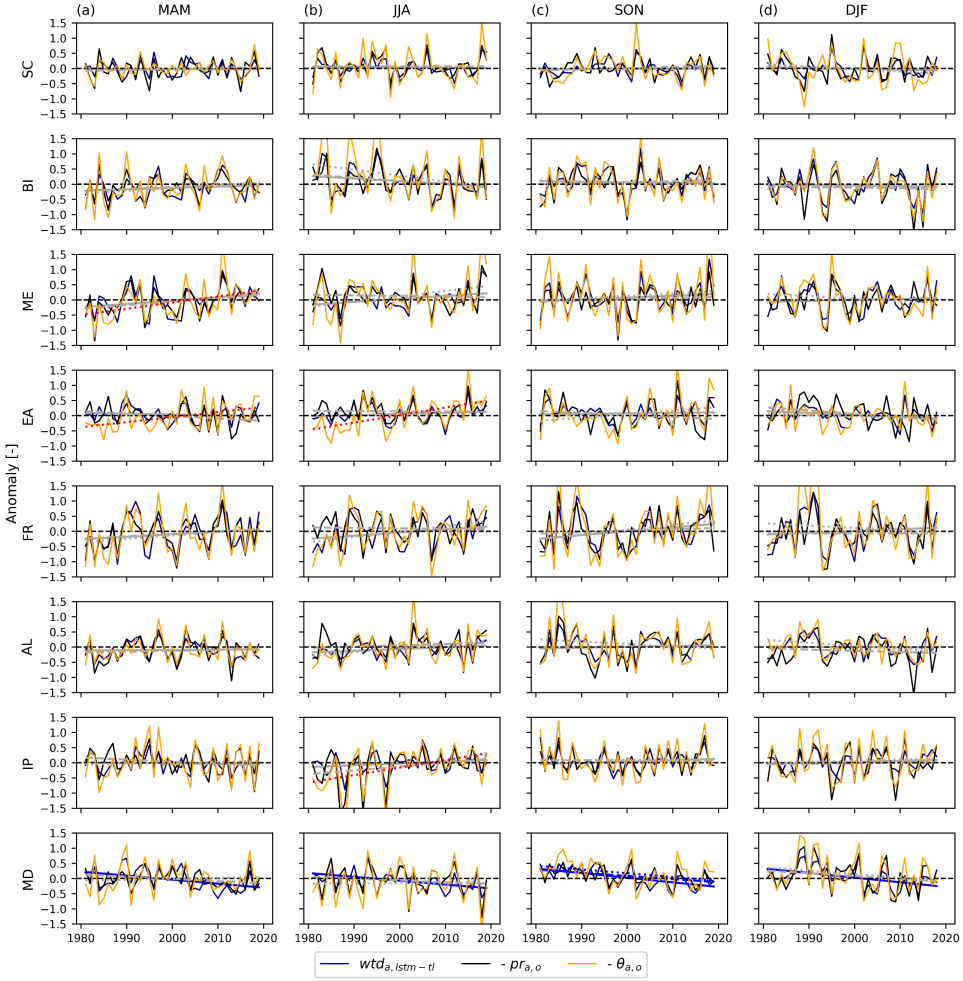

Figure C.3: Seasonal averaged ERA5 bias-corrected $pr_{a,o}$, ERA5-Land $\theta_{a,o}$ and RD1 $wtd_{a,lstm\text{-}tl}$ and their trends in eight PRUDENCE regions (MAM - March, April and May, i.e., spring; JJA - June, July and August, i.e., summer; SON - September, October and November, i.e., autumn; DJF - December, January and February, i.e., winter). The bold lines show the seasonal trends of - $pr_{a,o}$ (dashed), - $\theta_{a,o}$ (dotted) and $wtd_{a,lstm\text{-}tf}$ (solid), where red, blue and gray indicate a drier GW condition at the 95% significant level, a wetter GW condition at the 95% significant level and no significant trend, respectively. The black dashed lines indicate - $pr_{a,o}$/- $\theta_{a,o}$/wtd_a = 0. A long time period with - $pr_{a,o}$/- $\theta_{a,o}$/wtd_a > 0 corresponds to a dry event while a long time period with - $pr_{a,o}$/- $\theta_{a,o}$/wtd_a < 0 corresponds to a wet event.

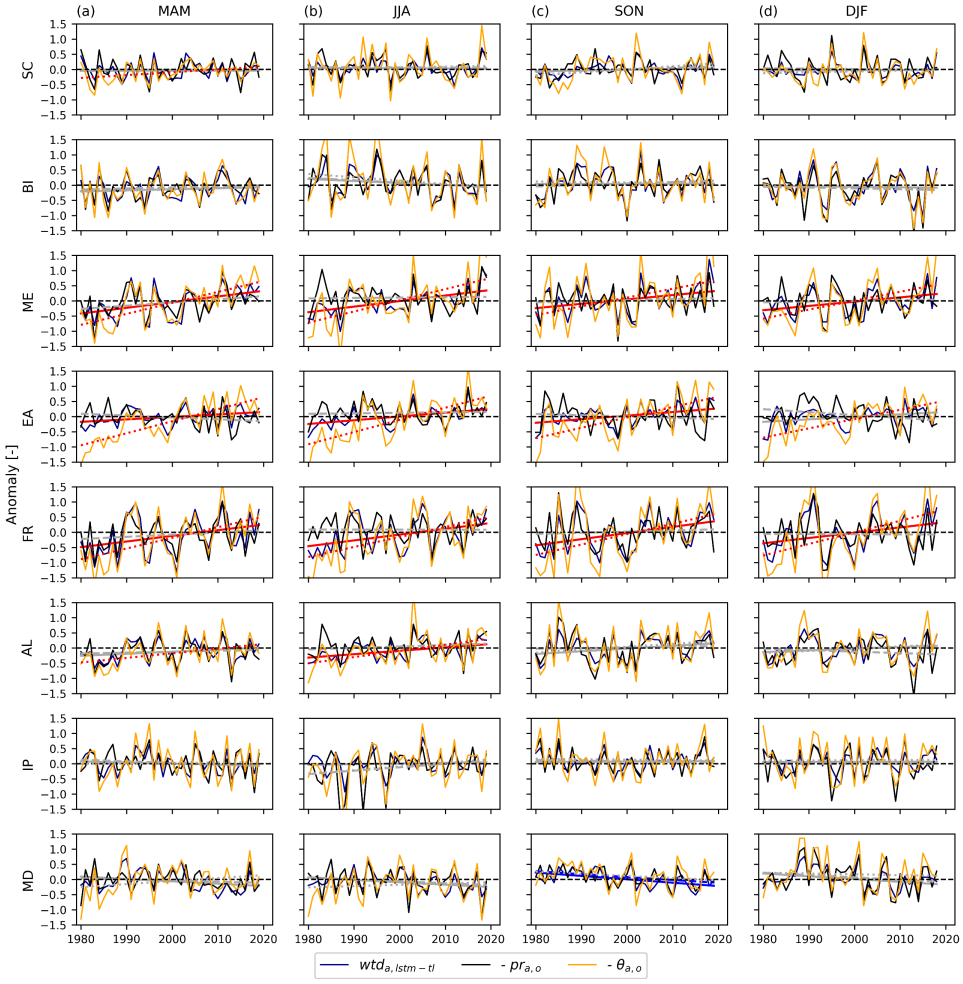

Figure C.4: Same as Figure C.3, but for ERA5 bias-corrected $pr_{a,o}$, GLEAM $\theta_{a,o}$ and RD2 $wtd_{a,lstm\text{-}tl}$.

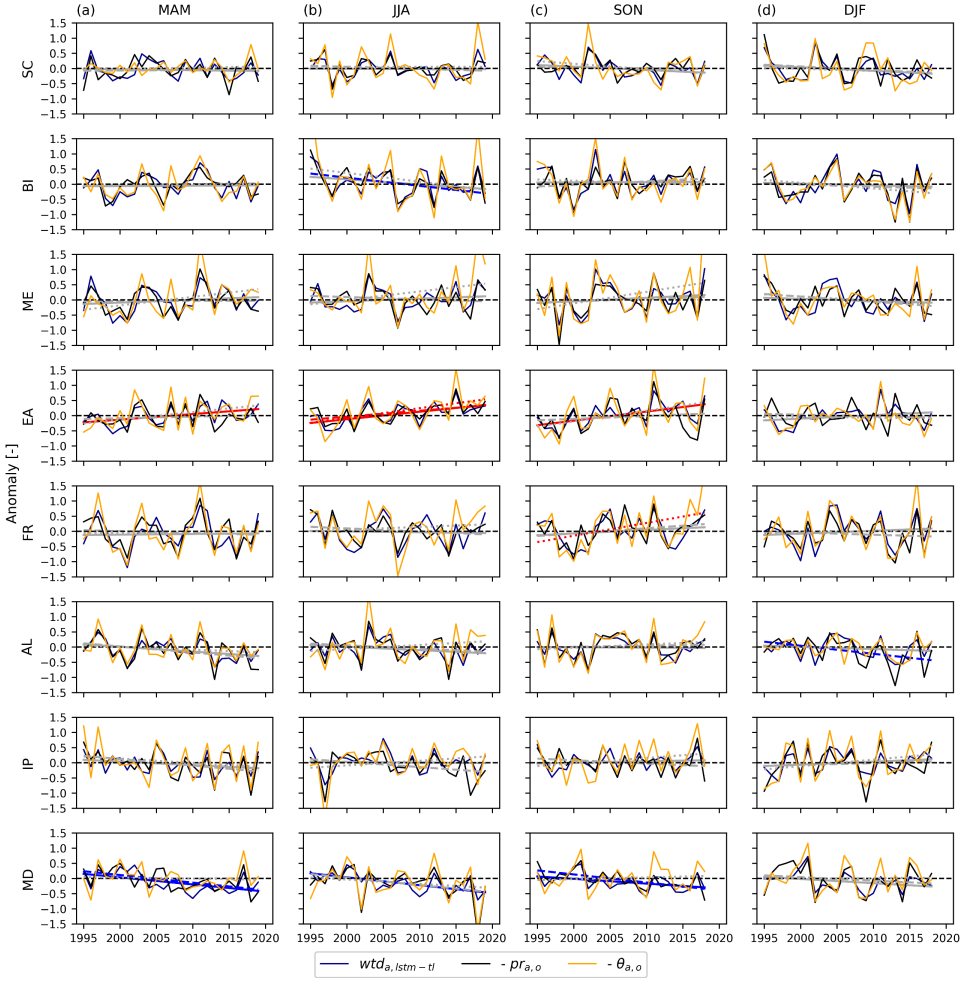

Figure C.5: Same as Figure C.3, but for COSMO-REA6 $pr_{a,o}$, ERA5-Land $\theta_{a,o}$ and RD3 $wtd_{a,lstm\text{-}tl}$.

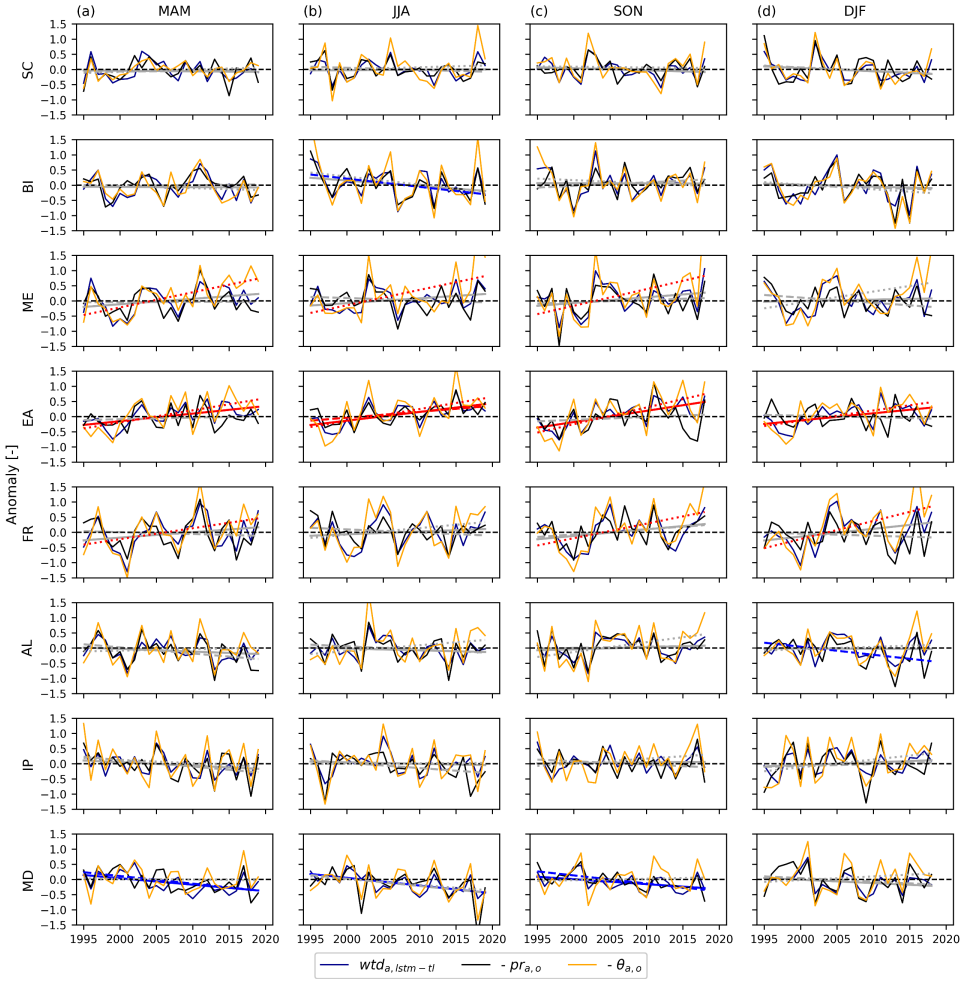

Figure C.6: Same as Figure C.3, but for COSMO-REA6 $pr_{a,o}$, GLEAM $\theta_{a,o}$ and RD4 $wtd_{a,lstm\text{-}tl}$.

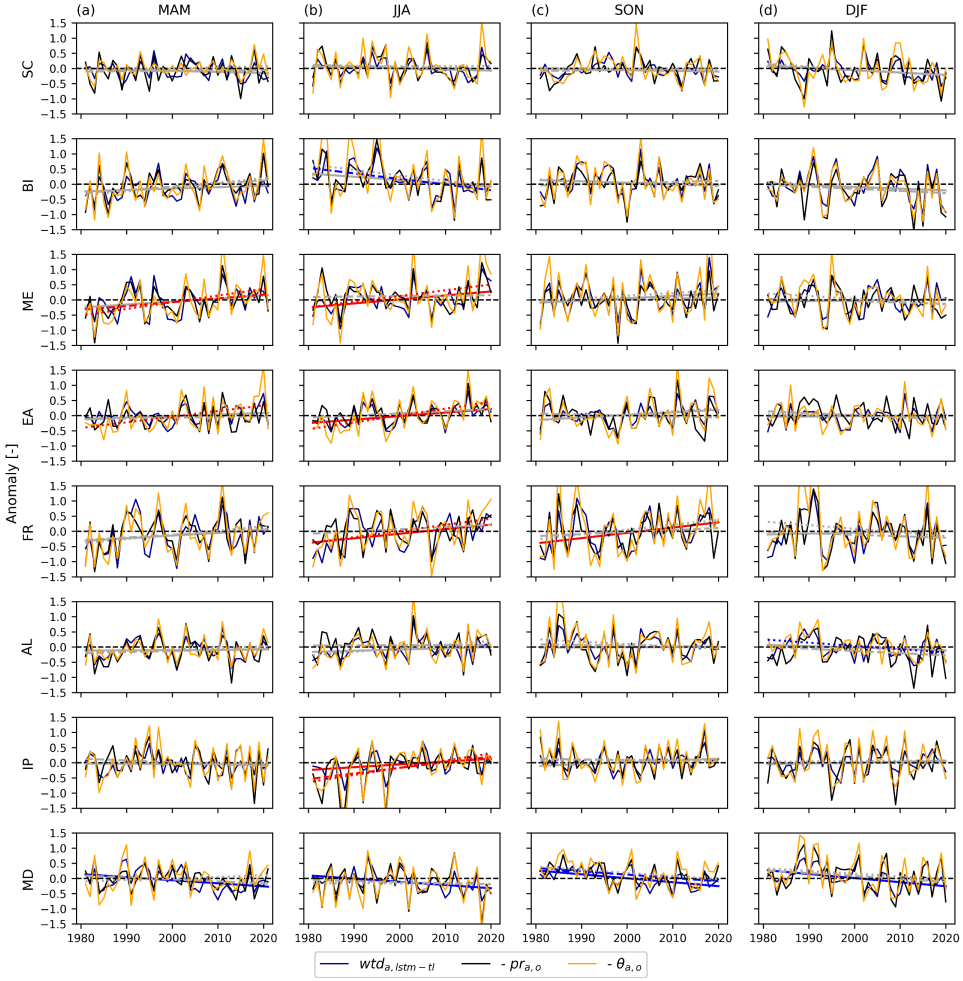

Figure C.7: Same as Figure C.3, but for ERA5-Land $pr_{a,o}$, ERA5-Land $\theta_{a,o}$ and RD5 $wtd_{a,lstm\text{-}tl}$.

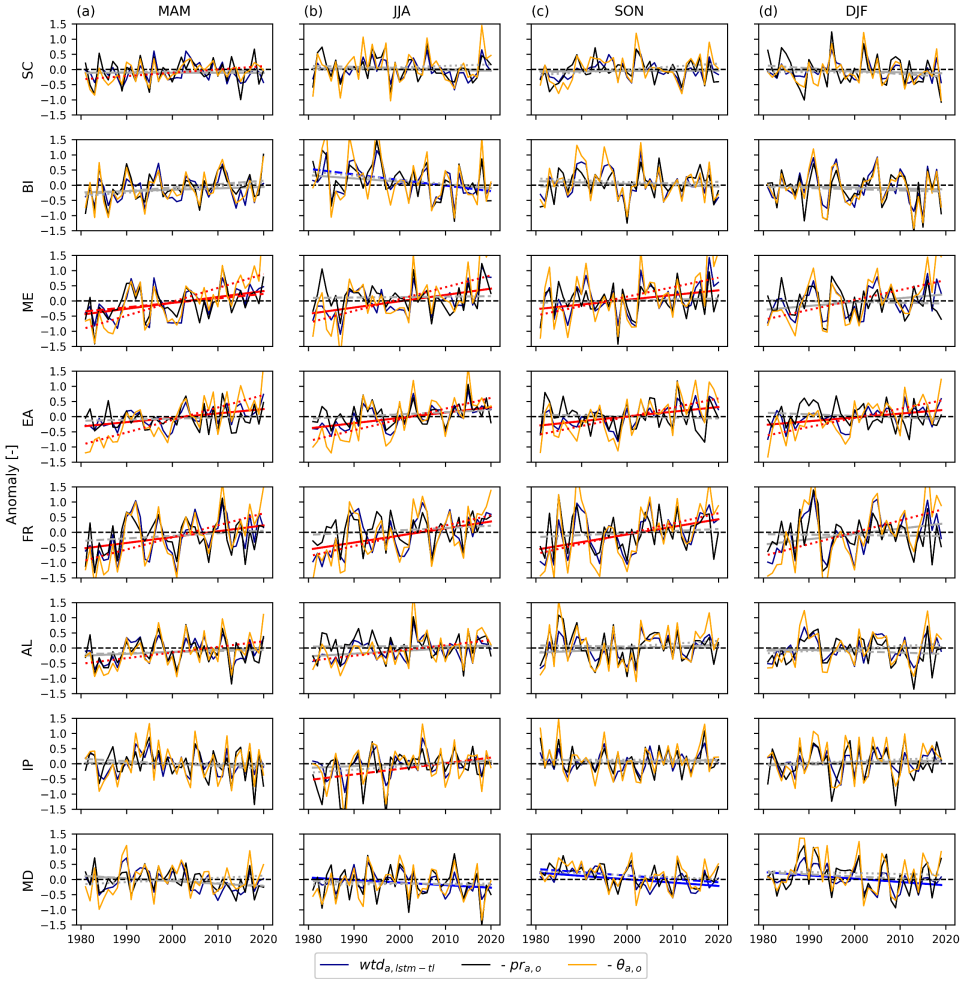

Figure C.8: Same as Figure C.3, but for ERA5-Land $pr_{a,o}$, GLEAM $\theta_{a,o}$ and RD6 $wtd_{a,lstm\text{-}tl}$.

Table C.1: Detailed information of European GW monitoring wells applied in Chapter 4.

Country/region	Source	Number of GW monitoring wells
France	ADES (https://ades.eaufrance.fr/Recherche/Index/Piezometre?g=6b2839, last access: June 2020)	91
The Netherlands	DINOloket (https://www.dinoloket.nl/en/subsurface-data, last access: June, 2020)	83
The UK	British Geological Survey	4
Sweden	Geological Survey of Sweden (SGU) (https://apps.sgu.se/kartvisare/kartvisare-grundvattenniva.html, last access: September, 2020)	38
Germany/Baden-Württemberg	Landesanstalt für Umwelt Baden-Württemberg (LUBW) (https://udo.lubw.baden-wuerttemberg.de/public/api/processingChain?ssid=324bd7d5-609d-4e8b-be1f-af7e073d56d5&selector=gwMessstellenauswahl.meros%3Ameros_z_gw_messwerte_uis_gwstand_messstellen_refdb%24ind1.sel, last access: July, 2020)	301
Germany/Bayern	Bayerisches Landesamt für Umwelt (https://www.gkd.bayern.de/en/groundwater/upper-layer/tables, last access: July, 2020)	87
Germany/Hessen	Hessisches Landesamt für Naturschutz, Umwelt und Geologie (http://lgd.hessen.de/mapapps/resources/apps/lgd/index.html?lang=en,last access: August, 2020)	368
Germany/Mecklenburg Vorpommern/ Mittleres Mecklenburg	Staatliches Amt für Landwirtschaft und Umwelt Mittleres Mecklenburg, Abteilung 4 Naturschutz, Wasser und Boden	8
Germany/Mecklenburg Vorpommern/Vorpommern	Staatliches Amt für Landwirtschaft und Umwelt Vorpommern, Abteilung 4 - Naturschutz, Wasser und Boden, Dezernat 44 - Wasserrahmenrichtlinie, Gewässerkunde	17
Germany/Rheinland-Pfalz	Landesamt für Umwelt Rheinland-Pfalz (http://www.gda-wasser.rlp.de/GDAWasser/client/gisclient/index.html?applicationId=12366&forcePreventCache=14143139175, last access: August, 2020)	210
Germany/Sachsen	Sächsisches Landesamt für Umwelt, Landwirtschaft und Geologie (https://www.umwelt.sachsen.de/umwelt/infosysteme/ida/pages/map/default/index.xhtml, last access: August, 2020)	380
Germany/Sachsen-Anhalt	Landesbetrieb für Hochwasserschutz und Wasserwirtschaft des Landes Sachsen-Anhalt (LHW) (http://www.lhw.sachsen-anhalt.de\gld-portal, last access: August, 2020)	397
Germany/Schleswig-Holstein	Landesamt für Landwirtschaft, Umwelt und ländliche Räume des Landes Schleswig-Holstein (http://www.umweltdaten.landsh.de/atlas/script/index.php, last access: August, 2020)	38
Germany/Hamburg	Behörde für Umwelt, Klima, Energie und Agrarwirtschaft (BUKEA)	1
Germany/Niedersachsen	Niedersächsischen Landesdatenbank für wasserwirtschaftliche Daten (http://www.wasserdaten.niedersachsen.de)	581

Table C.2: Numbers of pixels with $wtd_{a,o}$ in eight PRUDENCE regions.

SC	BI	ME	EA	FR	AL	IP	MD
12	3	592	0	27	41	4	2

Table C.3: Regional median r, RMSE, NSE and KGE of $wtd_{a,lstm(m)}$, $wtd_{a,m}$, $wtd_{a,lstm\text{-}tl}$ and $wtd_{a,lstm(o)}$ in seven PRUDENCE regions for the 1996-2016 period.

		SC	BI	ME	FR	AL	IP	MD
r [-]	$wtd_{a,lstm(m)}$	0.31	0.54	0.28	0.28	0.27	0.40	0.21
	$wtd_{a,m}$	0.24	0.52	0.26	0.25	0.26	0.33	0.21
	$wtd_{a,lstm\text{-}tl}$	0.57	0.63	0.44	0.44	0.45	0.45	0.37
	$wtd_{a,lstm(o)}$	0.86	0.78	0.89	0.84	0.89	0.83	0.49
RMSE [-]	$wtd_{a,lstm(m)}$	1.07	0.89	1.10	1.08	1.10	1.02	1.24
	$wtd_{a,m}$	1.28	0.97	1.22	1.29	1.23	1.21	1.39
	$wtd_{a,lstm\text{-}tl}$	0.88	0.77	0.97	0.97	0.95	0.98	1.15
	$wtd_{a,lstm(o)}$	0.51	0.61	0.48	0.60	0.49	0.60	1.00
NSE [-]	$wtd_{a,lstm(m)}$	-0.21	0.15	-0.18	-0.22	-0.18	-0.01	-0.23
	$wtd_{a,m}$	-0.61	-0.01	-0.46	-0.53	-0.45	-0.23	-0.54
	$wtd_{a,lstm\text{-}tl}$	0.18	0.35	0.08	0.05	0.12	0.12	-0.05
	$wtd_{a,lstm(o)}$	0.73	0.60	0.77	0.67	0.77	0.62	0.20
KGE [-]	$wtd_{a,lstm(m)}$	0.24	0.53	0.21	0.22	0.18	0.35	0.13
	$wtd_{a,m}$	0.23	0.52	0.25	0.25	0.25	0.33	0.20
	$wtd_{a,lstm\text{-}tl}$	0.43	0.60	0.34	0.36	0.34	0.37	0.29
	$wtd_{a,lstm(o)}$	0.76	0.62	0.77	0.69	0.77	0.59	0.16

List of Figures

List of Tables

Bibliography

Adamowski, J. and Chan, H. F.: A wavelet neural network conjunction model for groundwater level forecasting, J. Hydrol., 407(1–4), 28–40, doi:10.1016/j.jhydrol.2011.06.013, 2011.

Aeschbach-Hertig, W. and Gleeson, T.: Regional strategies for the accelerating global problem of groundwater depletion, Nat. Geosci., 5(12), 853–861, doi:10.1038/ngeo1617, 2012.

Aires, F., Weston, P., de Rosnay, P. and Fairbairn, D.: Statistical approaches to assimilate ASCAT soil moisture information—I. Methodologies and first assessment, Q. J. R. Meteorol. Soc., 147(736), 1823–1852, doi:10.1002/qj.3997, 2021.

Alley, W. M., Healy, R. W., LaBaugh, J. W. and Reilly, T. E.: Flow and Storage in Groundwater Systems, Science, 296(5575), 1985–1990, doi:10.1126/science.1067123, 2002.

Andersen, O. B., Seneviratne, S. I., Hinderer, J. and Viterbo, P.: GRACE-derived terrestrial water storage depletion associated with the 2003 European heat wave, Geophys. Res. Lett., 32(18), 1–4, doi:10.1029/2005GL023574, 2005.

Arras, L., Arjona-Medina, J., Widrich, M., Montavon, G., Gillhofer, M., Müller, K.-R., Hochreiter, S. and Samek, W.: Explaining and Interpreting LSTMs, pp. 211–238., 2019.

Ault, T. R.: On the essentials of drought in a changing climate, Science, 368(6488), 256–260, doi:10.1126/science.aaz5492, 2020.

Bachmair, S., Tanguy, M., Hannaford, J. and Stahl, K.: How well do meteorological indicators represent agricultural and forest drought across Europe?, Environ. Res. Lett., 13(3), doi:10.1088/1748-9326/aaafda, 2018.

Banerjee, S. and Mitra, M.: Application of cross wavelet transform for ECG pattern analysis and classification, IEEE Trans. Instrum. Meas., 63(2), 326–333, doi:10.1109/TIM.2013.2279001, 2014.

Bastos, A., Ciais, P., Friedlingstein, P., Sitch, S., Pongratz, J., Fan, L., Wigneron, J. P., Weber, U., Reichstein, M., Fu, Z., Anthoni, P., Arneth, A., Haverd, V., Jain, A. K., Joetzjer, E., Knauer, J., Lienert, S., Loughran, T., McGuire, P. C., Tian, H., Viovy, N. and Zaehle, S.: Direct and seasonal legacy effects of the 2018 heat wave and drought on European ecosystem productivity, Sci. Adv., 6(24), eaba2724, doi:10.1126/sciadv.aba2724, 2020.

Beck, H. E., Pan, M., Miralles, D. G., Reichle, R. H., Dorigo, W. A., Hahn, S., Sheffield, J., Karthikeyan, L., Balsamo, G., Parinussa, R. M., van Dijk, A. I. J. M., Du, J., Kimball, J. S., Vergopolan, N. and Wood, E. F.: Evaluation of 18 satellite- and model-based soil moisture products using in situ measurements from 826 sensors, Hydrol. Earth Syst. Sci., 25(1), 17–40, doi:10.5194/hess-25-17-2021, 2021.

Bloomfield, J. P. and Marchant, B. P.: Analysis of groundwater drought building on the standardised precipitation index approach, Hydrol. Earth Syst. Sci., 17(12), 4769–4787, doi:10.5194/hess-17-4769-2013, 2013.

Bloomfield, J. P., Marchant, B. P. and McKenzie, A. A.: Changes in groundwater drought associated with anthropogenic warming, Hydrol. Earth Syst. Sci., 23(3), 1393–1408, doi:10.5194/hess-23-1393-2019, 2019.

Bloomfield, J., Brauns, B., Hannah, D. M., Jackson, C., Marchant, B., Van Loon, A. F.: The Groundwater Drought Initiative (GDI): analysing and understanding groundwater drought across Europe, EGU General Assembly, Vienna, Austria, 8-13 April 2018, EGU2018-4540, 2018.

Boergens, E., Güntner, A., Dobslaw, H. and Dahle, C.: Quantifying the Central European Droughts in 2018 and 2019 With GRACE Follow-On, Geophys. Res. Lett., 47(14), doi:10.1029/2020GL087285, 2020.

Bollmeyer, C., Keller, J. D., Ohlwein, C., Wahl, S., Crewell, S., Friederichs, P., Hense, A., Keune, J., Kneifel, S., Pscheidt, I., Redl, S. and Steinke, S.: Towards a high-resolution regional reanalysis for the European CORDEX domain, Q. J. R. Meteorol. Soc., 141(686), 1–15, doi:10.1002/qj.2486, 2015.

Bonsor, H., Shamsudduha, M., Marchant, B., MacDonald, A. and Taylor, R.: Seasonal and Decadal Groundwater Changes in African Sedimentary Aquifers Estimated Using GRACE Products and LSMs, Remote Sens., 10(6), 904, doi:10.3390/rs10060904, 2018.

Brauns, B., Cuba, D., Bloomfield, J. P., Hannah, D. M., Jackson, C., Marchant, B. P., Heudorfer, B., Van Loon, A. F., Bessière, H., Thunholm, B. and Schubert, G.: The Groundwater Drought Initiative (GDI): Analysing and understanding groundwater drought across Europe, Proc. Int. Assoc. Hydrol. Sci., 383, 297–305, doi:10.5194/piahs-383-297-2020, 2020.

Brownlee, J.: How to Develop an Ensemble of Deep Learning Models in Keras, available at: https://machinelearningmastery.com/model-averaging-ensemble-for-deep-learning-neural-networks/, last access: November 2019, 2018.

Brutsaert, W.: THE HYDROLOGIC CYCLE, in Hydrology: An Introduction, pp. 2–3, Cambridge University Press, New York., 2005.

Chen, J., Famigliett, J. S., Scanlon, B. R. and Rodell, M.: Groundwater Storage Changes: Present Status from GRACE Observations, Surv. Geophys., 37(2), 397–417, doi:10.1007/s10712-015-9332-4, 2016.

Christensen, J. H. and Christensen, O. B.: A summary of the PRUDENCE model projections of changes in European climate by the end of this century, Clim. Change, 81(SUPPL. 1), 7–30, doi:10.1007/s10584-006-9210-7, 2007.

Cucchi, M., Weedon, G. P., Amici, A., Bellouin, N., Lange, S., Müller Schmied, H., Hersbach, H. and Buontempo, C.: WFDE5: bias-adjusted ERA5 reanalysis data for impact studies, Earth Syst. Sci. Data, 12(3), 2097–2120, doi:10.5194/essd-12-2097-2020, 2020.

Dawson, C. W. and Wilby, R. L.: Hydrological modelling using artificial neural networks, Prog. Phys. Geogr., 25(1), 80–108, doi:10.1177/030913330102500104, 2001.

de Graaf, I. E. M., Gleeson, T., (Rens) van Beek, L. P. H., Sutanudjaja, E. H. and Bierkens, M. F. P.: Environmental flow limits to global groundwater pumping, Nature, 574(7776), 90–94, doi:10.1038/s41586-019-1594-4, 2019.

Diersch, H.-J. G.: FEFLOW, Springer Berlin Heidelberg, Berlin, Heidelberg., 2014.

Dong, B., Sutton, R., Shaffrey, L. and Wilcox, L.: The 2015 European Heat Wave, Bull. Am. Meteorol. Soc., 97(12), S57–S62, doi:10.1175/BAMS-D-16-0140.1, 2016.

Duveiller, G., Fasbender, D. and Meroni, M.: Revisiting the concept of a symmetric index of agreement for continuous datasets, Sci. Rep., 6(October 2015), 1–14, doi:10.1038/srep19401, 2016.

EEA: Amount of groundwater abstraction, in Groundwater quality and quantity in Europe, Office for Official Publications of the European Communities, Luxembourg., 1999.

EEA: Meteorological and hydrological droughts in Europe, [online] Available from: https://www.eea.europa.eu/data-and-maps/indicators/river-flow-drought-3/assessment (Accessed 7 July 2020), 2020.

EEA: Meteorological and hydrological droughts, available at: https://www.eea.europa.eu/data-and-maps/indicators/river-flow-drought-2/assessment, last access: January 2019, 2016.

Entekhabi, D., Njoku, E. G., O'Neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., Entin, J. K., Goodman, S. D., Jackson, T. J., Johnson, J., Kimball, J., Piepmeier, J. R., Koster, R. D., Martin, N., McDonald, K. C., Moghaddam, M., Moran, S., Reichle, R., Shi, J. C., Spencer, M. W., Thurman, S. W., Tsang, L. and Van Zyl, J.: The Soil Moisture Active Passive (SMAP) Mission, Proc. IEEE, 98(5), 704–716, doi:10.1109/JPROC.2010.2043918, 2010.

European Commission: Groundwater Protection in Europe: THE NEW GROUNDWATER DIRECTIVE – CONSOLIDATING THE EU REGULATORY FRAMEWORK, Brussels., 2008.

Fan, Y., Li, H. and Miguez-Macho, G.: Global Patterns of Groundwater Table Depth, Science, 339(6122), 940–943, doi:10.1126/science.1229881, 2013.

Fang, Z., Bogena, H., Kollet, S., Koch, J. and Vereecken, H.: Spatio-temporal validation of long-term 3D hydrological simulations of a forested catchment using empirical orthogonal functions and wavelet coherence analysis, J. Hydrol., 529, 1754–1767, doi:10.1016/j.jhydrol.2015.08.011, 2015.

Fink, A. H., Brücher, T., Krüger, A., Leckebusch, G. C., Pinto, J. G. and Ulbrich, U.: The 2003 European summer heatwaves and drought -synoptic diagnosis and impacts, Weather, 59(8), 209–216, doi:10.1256/wea.73.04, 2004.

Fiorillo, F. and Guadagno, F. M.: Long karst spring discharge time series and droughts occurrence in Southern Italy, Environ. Earth Sci., 65(8), 2273–2283, doi:10.1007/s12665-011-1495-9, 2012.

Freeze, R. A. and Witherspoon, P. A.: Theoretical analysis of regional groundwater flow: 1. Analytical and numerical solutions to the mathematical model, Water Resour. Res., 2(4), 641–656, doi:10.1029/WR002i004p00641, 1966.

Furusho-Percot, C., Goergen, K., Hartick, C., Kulkarni, K., Keune, J. and Kollet, S.: Pan-European groundwater to atmosphere terrestrial systems climatology from a physically consistent simulation, Sci. data, 6(1), 320, doi:10.1038/s41597-019-0328-7, 2019.

Gasper, F., Goergen, K., Shrestha, P., Sulis, M., Rihani, J., Geimer, M. and Kollet, S.: Implementation and scaling of the fully coupled Terrestrial Systems Modeling Platform (TerrSysMP v1.0) in a massively parallel supercomputing environment - A case study on JUQUEEN (IBM Blue Gene/Q), Geosci. Model Dev., 7(5), 2531–2543, doi:10.5194/gmd-7-2531-2014, 2014.

Gers, F. A., Schmidhuber, J. and Cummins, F.: Learning to Forget: Continual Prediction with LSTM, Neural Comput., 12(10), 2451–2471, doi:10.1162/089976600300015015, 2000.

Gholami, V., Chau, K. W., Fadaee, F., Torkaman, J. and Ghaffari, A.: Modeling of groundwater level fluctuations using dendrochronology in alluvial aquifers, J. Hydrol., 529, 1060–1069, doi:10.1016/j.jhydrol.2015.09.028, 2015.

Gong, Y., Zhang, Y., Lan, S. and Wang, H.: A Comparative Study of Artificial Neural Networks, Support Vector Machines and Adaptive Neuro Fuzzy Inference System for Forecasting Groundwater Levels near Lake Okeechobee, Florida, Water Resour. Manag., 30(1), 375–391, doi:10.1007/s11269-015-1167-8, 2016.

Goodfellow, I., Bengio, Y. and Courville, A.: Introduction, in Deep learning, pp. 3–8, MIT Press, Cambridge, Massachusetts, 2017a.

Goodfellow, I., Bengio, Y. and Courville, A.: Bagging and Other Ensemble Methods, pp. 250–251, MIT Press, Cambridge, Massachusetts, 2017b.

Goodfellow, I., Bengio, Y. and Courville, A.: Transfer Learning and Domain Adaptation, in Deep learning, pp. 526–531, MIT Press, Cambridge, Massachusetts, 2017c.

Govindaraju, R.: Artificial Neural Networks in Hydrology. I: Preliminary Concepts, J. Hydrol. Eng., 5(2), 115–123, doi:10.1061/(ASCE)1084-0699(2000)5:2(115), 2000.

Green, T. R., Taniguchi, M., Kooi, H., Gurdak, J. J., Allen, D. M., Hiscock, K. M., Treidel, H. and Aureli, A.: Beneath the surface of global change: Impacts of climate change on groundwater, J. Hydrol., 405(3–4), 532–560, doi:10.1016/j.jhydrol.2011.05.002, 2011.

Grinsted, A., Moore, J. C. and Jevrejeva, S.: Application of the cross wavelet transform and wavelet coherence to geophysical time series, Nonlinear Process. Geophys., 11(5/6), 561–566, doi:10.5194/npg-11-561-2004, 2004.

Gruber, A., Scanlon, T., Van Der Schalie, R., Wagner, W. and Dorigo, W.: Evolution of the ESA CCI Soil Moisture climate data records and their underlying merging methodology, Earth Syst. Sci. Data, 11(2), 717–739, doi:10.5194/essd-11-717-2019, 2019.

Guerreiro, S. B., Dawson, R. J., Kilsby, C., Lewis, E. and Ford, A.: Future heat-waves, droughts and floods in 571 European cities, Environ. Res. Lett., 13(3), 034009, doi:10.1088/1748-9326/aaaad3, 2018.

Han, Z., Huang, S., Huang, Q., Leng, G., Wang, H., Bai, Q., Zhao, J., Ma, L., Wang, L. and Du, M.: Propagation dynamics from meteorological to groundwater drought and their possible influence factors, J. Hydrol., 578, 124102, doi:10.1016/j.jhydrol.2019.124102, 2019.

Hänsel, S., Ustrnul, Z., Łupikasza, E. and Skalak, P.: Assessing seasonal drought variations and trends over Central Europe, Adv. Water Resour., 127(September 2018), 53–75, doi:10.1016/j.advwatres.2019.03.005, 2019.

Hari, V., Rakovec, O., Markonis, Y., Hanel, M. and Kumar, R.: Increased future occurrences of the exceptional 2018–2019 Central European drought under global warming, Sci. Rep., 10(1), 12207, doi:10.1038/s41598-020-68872-9, 2020.

Hartick, C., Furusho-Percot, C., Goergen, K. and Kollet, S.: An Interannual Probabilistic Assessment of Subsurface Water Storage Over Europe Using a Fully Coupled Terrestrial Model, Water Resour. Res., 57(1), doi:10.1029/2020WR027828, 2021.

Hauck, W. W. and Donner, A.: Wald's Test as Applied to Hypotheses in Logit Analysis, J. Am. Stat. Assoc., 72(360a), 851–853, doi:10.1080/01621459.1977.10479969, 1977.

Hauswirth, S. M., Bierkens, M. F. P., Beijk, V. and Wanders, N.: The potential of data driven approaches for quantifying hydrological extremes, Adv. Water Resour., 155, 104017, doi:10.1016/j.advwatres.2021.104017, 2021.

Haykin, S.: WHAT IS A NEURAL NETWORK?, in Neural Networks and Learning Machines, pp. 1–2, Prentice Hall, New York., 2009.

Hellwig, J., Graaf, I. E. M., Weiler, M. and Stahl, K.: Large-Scale Assessment of Delayed Groundwater Responses to Drought, Water Resour. Res., 56(2), doi:10.1029/2019WR025441, 2020.

Hinton, G., Srivastava, N. and Swersky, K.: Overview of mini-batch gradient descent, available at: https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf, last access: January 2020, 2012.

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Comput., 9(8), 1735–1780, doi:10.1162/neco.1997.9.8.1735, 1997.

Hou, A. Y., Kakar, R. K., Neeck, S., Azarbarzin, A. A., Kummerow, C. D., Kojima, M., Oki, R., Nakamura, K. and Iguchi, T.: The Global Precipitation Measurement Mission, Bull. Am. Meteorol. Soc., 95(5), 701–722, doi:10.1175/BAMS-D-13-00164.1, 2014.

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., Hong, Y., Bowman, K. P. and Stocker, E. F.: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales, J. Hydrometeorol., 8(1), 38–55, doi:10.1175/JHM560.1, 2007.

Hughes, J. D., Langevin, C. D. and Banta, E. R.: Documentation for the MODFLOW 6 Framework, in U.S. Geological Survey Techniques and Methods 6–A57, edited by W. H. Werkheiser, pp. 1–30, U.S. Geological Survey, Reston, VA., 2017.

Jasechko, S. and Perrone, D.: Global groundwater wells at risk of running dry, Science, 372(6540), 418–421, doi:10.1126/science.abc2755, 2021.

Jeong, J., Park, E., Chen, H., Kim, K.-Y., Shik Han, W. and Suk, H.: Estimation of groundwater level based on the robust training of recurrent neural networks using corrupted data, J. Hydrol., 582, 124512, doi:10.1016/j.jhydrol.2019.124512, 2020.

Jones, P. W.: First- and Second-Order Conservative Remapping Schemes for Grids in Spherical Coordinates, Mon. Weather Rev., 127(9), 2204–2210, doi:10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2, 1999.

JSC: JUWELS – Installation of JUQUEEN Successor Started, [online] Available from: https://www.fz-juelich.de/SharedDocs/Meldungen/IAS/JSC/EN/2018/2018-01-juwels.html?nn=2274286 (Accessed 25 October 2021), 2018.

Karim, M. N. and Rivera, S. L.: Comparison of feed-forward and recurrent neural networks for bioprocess state estimation, Comput. Chem. Eng., 16, S369–S377, doi:10.1016/S0098-1354(09)80044-6, 1992.

Kenda, K., Čerin, M., Bogataj, M., Senožetnik, M., Klemen, K., Pergar, P., Laspidou, C. and Mladenić, D.: Groundwater Modeling with Machine Learning Techniques: Ljubljana polje Aquifer, Proceedings, 2(11), 697, doi:10.3390/proceedings2110697, 2018.

Kerr, Y. H., Waldteufel, P., Wigneron, J.-P., Delwart, S., Cabot, F., Boutin, J., Escorihuela, M.-J., Font, J., Reul, N., Gruhier, C., Juglea, S. E., Drinkwater, M. R., Hahne, A., Martín-Neira, M. and Mecklenburg, S.: The SMOS Mission: New Tool for Monitoring Key Elements ofthe Global Water Cycle, Proc. IEEE, 98(5), 666–687, doi:10.1109/JPROC.2010.2043032, 2010.

Keune, J., Gasper, F., Goergen, K., Hense, A., Shrestha, P., Sulis, M. and Kollet, S.: Studying the influence of groundwater representations on land surface-atmosphere feedbacks during the European heat wave in 2003, J. Geophys. Res., 121(22), 13,301-13,325, doi:10.1002/2016JD025426, 2016.

Keune, J., Sulis, M. and Kollet, S. J.: Potential Added Value of Incorporating Human Water Use on the Simulation of Evapotranspiration and Precipitation in a Continental-Scale Bedrock-to-Atmosphere Modeling System: A Validation Study Considering Observational Uncertainty, J. Adv. Model. Earth Syst., 11(7), 1959–1980, doi:10.1029/2019MS001657, 2019.

Khan, S., Gabriel, H. F. and Rana, T.: Standard precipitation index to track drought and assess impact of rainfall on watertables in irrigation areas, Irrig. Drain. Syst., 22(2), 159–177, doi:10.1007/s10795-008-9049-3, 2008.

Knoben, W. J. M., Freer, J. E. and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrol. Earth Syst. Sci., 23(10), 4323–4331, doi:10.5194/hess-23-4323-2019, 2019.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K. and Herrnegger, M.: Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks, Hydrol. Earth Syst. Sci., 22(11), 6005–6022, doi:10.5194/hess-22-6005-2018, 2018.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S. and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrol. Earth Syst. Sci., 23(12), 5089–5110, doi:10.5194/hess-23-5089-2019, 2019.

Kumar, R., Musuuza, J. L., Van Loon, A. F., Teuling, A. J., Barthel, R., Ten Broek, J., Mai, J., Samaniego, L. and Attinger, S.: Multiscale evaluation of the Standardized Precipitation Index as a groundwater drought indicator, Hydrol. Earth Syst. Sci., 20(3), 1117–1131, doi:10.5194/hess-20-1117-2016, 2016.

Kurtz, W., He, G., Kollet, S. J., Maxwell, R. M., Vereecken, H. and Franssen, H. J. H.: TerrSysMP-PDAF (version 1.0): A modular high-performance data assimilation framework for an integrated land surface-subsurface model, Geosci. Model Dev., 9(4), 1341–1360, doi:10.5194/gmd-9-1341-2016, 2016.

Labat, D.: Recent advances in wavelet analyses: Part 1. A review of concepts, J. Hydrol., 314(1–4), 275–288, doi:10.1016/j.jhydrol.2005.04.003, 2005.

Lane, S. N.: Assessment of rainfall-runoff models based upon wavelet analysis, Hydrol. Process., 21(5), 586–607, doi:10.1002/hyp.6249, 2007.

Le, X. H., Ho, H. V., Lee, G. and Jung, S.: Application of Long Short-Term Memory (LSTM) neural network for flood forecasting, Water (Switzerland), 11(7), doi:10.3390/w11071387, 2019.

Lee, K.-H., He, X., Zhang, L. and Yang, L.: CleanNet: Transfer Learning for Scalable Image Classifier Training with Label Noise, [online] Available from: http://arxiv.org/abs/1711.07131, 2017.

Liong, S.-Y., Lim, W.-H. and Paudyal, G. N.: River Stage Forecasting in Bangladesh: Neural Network Approach, J. Comput. Civ. Eng., 14(1), 1–8, doi:10.1061/(ASCE)0887-3801(2000)14:1(1), 2000.

Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S. and Zhang, G.: Transfer learning using computational intelligence: A survey, Knowledge-Based Syst., 80, 14–23, doi:10.1016/j.knosys.2015.01.010, 2015.

Ma, K., Feng, D., Lawson, K., Tsai, W., Liang, C., Huang, X., Sharma, A. and Shen, C.: Transferring Hydrologic Data Across Continents – Leveraging Data-Rich Regions to Improve Hydrologic Prediction in Data-Sparse Regions, Water Resour. Res., 57(5), doi:10.1029/2020WR028600, 2021.

Ma, Y., Matta, E., Meißner, D., Schellenberg, H. and Hinkelmann, R.: Can machine learning improve the accuracy of water level forecasts for inland navigation? Case study: Rhine River Basin, Germany, 38th IAHR World Congr. Panama City 2019, Water - Connect. world, doi:10.3850/38WC092019-0274, 2019.

Malakar, P., Sarkar, S., Mukherjee, A., Bhanja, S. and Sun, A. Y.: Use of machine learning and deep learning methods in groundwater, in Global Groundwater, pp. 545–557, Elsevier., 2021.

Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A. and Verhoest, N. E. C.: GLEAM v3: satellite-based land evaporation and root-zone soil moisture, Geosci. Model Dev., 10(5), 1903–1925, doi:10.5194/gmd-10-1903-2017, 2017.

Maxwell, R. M., Kollet, S. J., Smith, S. G., Woodward, C. S., Falgout, R. D., Ferguson, I. M., Engdahl, N. B., Condon, L. E., Hector, B., Lopez, S., Gilbert, J., Bearup, L., Jefferson, J., Collins, C., De Graaf, I., Pribulick, C., Baldwin, C., Bosl, W. J., Hornung, R. and and Ashby, S.: ParFlow User's Manual, Integrated GroundWater Modeling Center Report, 2019.

Maxwell, R. M.: Infiltration in Arid Environments: Spatial Patterns between Subsurface Heterogeneity and Water-Energy Balances, Vadose Zo. J., 9(4), 970–983, doi:10.2136/vzj2010.0014, 2010.

Miralles, D. G., Holmes, T. R. H., De Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A. and Dolman, A. J.: Global land-surface evaporation estimated from satellite-based observations, Hydrol. Earth Syst. Sci., 15(2), 453–469, doi:10.5194/hess-15-453-2011, 2011.

McKee, T. B., Doesken, N. J. and Kleist, J.: THE RELATIONSHIP OF DROUGHT FREQUENCY AND DURATION TO TIME SCALES, in Proceedings of the 8th Conference on Applied Climatology, pp. 179–184, American Meteorological Society, Anaheim., 1993.

Mishra, A. K. and Singh, V. P.: A review of drought concepts, J. Hydrol., 391(1–2), 202–216, doi:10.1016/j.jhydrol.2010.07.012, 2010.

Mohanty, S., Jha, M. K., Raul, S. K., Panda, R. K. and Sudheer, K. P.: Using Artificial Neural Network Approach for Simultaneous Forecasting of Weekly Groundwater Levels at Multiple Sites, Water Resour. Manag., 29(15), 5521–5532, doi:10.1007/s11269-015-1132-6, 2015.

Müller, A. C. and Guido, S.: Introduction, in Introduction to machine learning with Python: A GUIDE FOR DATA SCIENTISTS, p. 1, O'Reilly Media, Inc., Sebastopol., 2017a.

Müller, A. C. and Guido, S.: Overfitting, and Underfitting, in Introduction to machine learning with Python: A GUIDE FOR DATA SCIENTISTS, p. 28-30, O'Reilly Media, Inc., Sebastopol., 2017b.

Müller, J., Park, J., Sahu, R., Varadharajan, C., Arora, B., Faybishenko, B. and Agarwal, D.: Surrogate optimization of deep neural networks for groundwater predictions, J. Glob. Optim., 81(1), 203–231, doi:10.1007/s10898-020-00912-0, 2021.

Muñoz Sabater, J.: Near surface meteorological variables from 1979 to 2019 derived from bias-corrected reanalysis, Copernicus Clim. Chang. Serv. Clim. Data Store, doi:10.24381/cds.20d54e34, 2021a.

Muñoz Sabater, J.: ERA5-Land hourly data from 1981 to present, Copernicus Clim. Chang. Serv. Clim. Data Store, doi:10.24381/cds.e2161bac, 2021b.

Murdoch, W. J., Liu, P. J. and Yu, B.: Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs, [online] Available from: http://arxiv.org/abs/1801.05453, 2018.

Naghibi, S. A., Pourghasemi, H. R. and Dixon, B.: 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(1), 1–27, doi:10.1007/s10661-015-5049-6, 2016.

Nalbantis, I. and Tsakiris, G.: Assessment of hydrological drought revisited, Water Resour. Manag., 23(5), 881–897, doi:10.1007/s11269-008-9305-1, 2009.

Nayak, P. C., Satyaji Rao, Y. R. and Sudheer, K. P.: Groundwater level forecasting in a shallow aquifer using artificial neural network approach, Water Resour. Manag., 20(1), 77–90, doi:10.1007/s11269-006-4007-z, 2006.

NOAA National Weather Service: The Hydrologic Cycle, [online] Available from: https://www.weather.gov/jetstream/hydro, 2015.

Norris, B.: July drought in Europe to cost at least €3.5bn, Aon - Commercial Risk, available at: https://www.commercialriskonline.com/july-drought-to-cost-at-least-e3-5bn-aon/, last access: January 2019, 2018.

Olah, C.: Understanding LSTM Networks -- colah's blog, available at: http://colah.github.io/posts/2015-08-Understanding-LSTMs/, last access: June 2018, 2015.

Palmer, W. C.: Keeping Track of Crop Moisture Conditions, Nationwide: The New Crop Moisture Index, Weatherwise, 21(4), 156–161, doi:10.1080/00431672.1968.9932814, 1968.

Perlman, H.: Where is Earth's water? USGS Water-Science School, available at: https://web.archive.org/web/20131214091601/http://ga.water.usgs.gov/edu/earthwherewater.html, last access: August 2019, 2013.

Prokoph, A. and El Bilali, H.: Cross-wavelet analysis: A tool for detection of relationships between paleoclimate proxy records, Math. Geosci., 40(5), 575–586, doi:10.1007/s11004-008-9170-8, 2008.

Quattoni, A., Collins, M. and Darrell, T.: Transfer learning for image classification with sparse prototype representations, in 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8, IEEE., 2008.

Rahmati, M., Groh, J., Graf, A., Pütz, T., Vanderborght, J. and Vereecken, H.: On the impact of increasing drought on the relationship between soil water content and evapotranspiration of a grassland, Vadose Zo. J., 19(1), doi:10.1002/vzj2.20029, 2020.

Reager, J. T., Gardner, A. S., Famiglietti, J. S., Wiese, D. N., Eicker, A. and Lo, M.-H.: A decade of sea level rise slowed by climate-driven hydrology, Science, 351(6274), 699–703, doi:10.1126/science.aad8386, 2016.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N. and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566(7743), 195–204, doi:10.1038/s41586-019-0912-1, 2019.

Rodell, M., Chen, J., Kato, H., Famiglietti, J. S., Nigro, J. and Wilson, C. R.: Estimating groundwater storage changes in the Mississippi River basin (USA) using GRACE, Hydrogeol. J., 15(1), 159–166, doi:10.1007/s10040-006-0103-7, 2007.

Rodell, M., Famiglietti, J. S., Wiese, D. N., Reager, J. T., Beaudoing, H. K., Landerer, F. W. and Lo, M.-H.: Emerging trends in global freshwater availability, Nature, 557(7707), 651–659, doi:10.1038/s41586-018-0123-1, 2018.

Sahoo, B. B., Jha, R., Singh, A. and Kumar, D.: Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting, Acta Geophys., 67(5), 1471–1481, doi:10.1007/s11600-019-00330-1, 2019.

Salerno, F. and Tartari, G.: A coupled approach of surface hydrological modelling and Wavelet Analysis for understanding the baseflow components of river discharge in karst environments, J. Hydrol., 376(1–2), 295–306, doi:10.1016/j.jhydrol.2009.07.042, 2009.

Shen, C.: A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists, Water Resour. Res., 54(11), 8558–8593, doi:10.1029/2018WR022643, 2018.

Shrestha, P., Sulis, M., Masbou, M., Kollet, S. and Simmer, C.: A Scale-Consistent Terrestrial Systems Modeling Platform Based on COSMO, CLM, and ParFlow, Mon. Weather Rev., 142(9), 3466–3483, doi:10.1175/MWR-D-14-00029.1, 2014.

Shukla, S. and Wood, A. W.: Use of a standardized runoff index for characterizing hydrologic drought, Geophys. Res. Lett., 35(2), 1–7, doi:10.1029/2007GL032487, 2008.

Simmer, C., Thiele-Eich, I., Masbou, M., Amelung, W., Bogena, H., Crewell, S., Diekkrüger, B., Ewert, F., Hendricks Franssen, H.-J., Huisman, J. A., Kemna, A., Klitzsch, N., Kollet, S., Langensiepen, M., Löhnert, U., Rahman, A. S. M. M., Rascher, U., Schneider, K., Schween, J., Shao, Y., Shrestha, P., Stiebler, M., Sulis, M., Vanderborght, J., Vereecken, H., van der Kruk, J., Waldhoff, G. and Zerenner, T.: Monitoring and Modeling the Terrestrial System from Pores to Catchments: The Transregional Collaborative Research Center on Patterns in the Soil–Vegetation–Atmosphere System, Bull. Am. Meteorol. Soc., 96(10), 1765–1787, doi:10.1175/BAMS-D-13-00134.1, 2015.

Sophocleous, M. A., Koelliker, J. K., Govindaraju, R. S., Birdie, T., Ramireddygari, S. R. and Perkins, S. P.: Integrated numerical modeling for basin-wide water management: The case of the Rattlesnake Creek basin in south-central Kansas, J. Hydrol., 214(1–4), 179–196, doi:10.1016/S0022-1694(98)00289-3, 1999.

Stagge, J. H., Kingston, D. G., Tallaksen, L. M. and Hannah, D. M.: Observed drought indices show increasing divergence across Europe, Sci. Rep., 7(1), 1–10, doi:10.1038/s41598-017-14283-2, 2017.

Stahl, K., Kohn, I., Blauhut, V., Urquijo, J., De Stefano, L., Acácio, V., Dias, S., Stagge, J. H., Tallaksen, L. M., Kampragou, E., Van Loon, A. F., Barker, L. J., Melsen, L. A., Bifulco, C., Musolino, D., De Carli, A., Massarutto, A., Assimacopoulos, D. and Van Lanen, H. A. J.: Impacts of European drought events: Insights from an international database of text-based reports, Nat. Hazards Earth Syst. Sci., 16(3), 801–819, doi:10.5194/nhess-16-801-2016, 2016.

Sulis, M., Keune, J., Shrestha, P., Simmer, C. and Kollet, S. J.: Quantifying the Impact of Subsurface-Land Surface Physical Processes on the Predictive Skill of Subseasonal Mesoscale Atmospheric Simulations, J. Geophys. Res. Atmos., 123(17), 9131–9151, doi:10.1029/2017JD028187, 2018.

Sun, A. Y. and Scanlon, B. R.: How can Big Data and machine learning benefit environment and water management: A survey of methods, applications, and future directions, Environ. Res. Lett., 14(7), doi:10.1088/1748-9326/ab1b7d, 2019.

Sun, A. Y.: Predicting groundwater level changes using GRACE data, Water Resour. Res., 49(9), 5900–5912, doi:10.1002/wrcr.20421, 2013.

Sun, Y., Wendi, D., Kim, D. E. and Liong, S. Y.: Technical note: Application of artificial neural networks in groundwater table forecasting-a case study in a Singapore swamp forest, Hydrol. Earth Syst. Sci., 20(4), 1405–1412, doi:10.5194/hess-20-1405-2016, 2016.

Sundararajan, M., Taly, A. and Yan, Q.: Axiomatic Attribution for Deep Networks, [online] Available from: http://arxiv.org/abs/1703.01365, 2017.

Sung, J. Y., Lee, J., Chung, I. M. and Heo, J. H.: Hourly water level forecasting at tributary affecteby main river condition, Water (Switzerland), 9(9), 1–17, doi:10.3390/w9090644, 2017.

Supreetha, B. S., Shenoy, N. and Nayak, P.: Lion Algorithm-Optimized Long Short-Term Memory Network for Groundwater Level Forecasting in Udupi District, India, Appl. Comput. Intell. Soft Comput., 2020, doi:10.1155/2020/8685724, 2020.

Tallaksen, L. M., Hisdal, H. and Lanen, H. A. J. Van: Space–time modelling of catchment scale drought characteristics, J. Hydrol., 375(3–4), 363–372, doi:10.1016/j.jhydrol.2009.06.032, 2009.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C. and Liu, C.: A Survey on Deep Transfer Learning, [online] Available from: http://arxiv.org/abs/1808.01974, 2018.

Tapley, B. D., Watkins, M. M., Flechtner, F., Reigber, C., Bettadpur, S., Rodell, M., Sasgen, I., Famiglietti, J. S., Landerer, F. W., Chambers, D. P., Reager, J. T., Gardner, A. S., Save, H., Ivins, E. R., Swenson, S. C., Boening, C., Dahle, C., Wiese, D. N., Dobslaw, H., Tamisiea, M. E. and Velicogna, I.: Contributions of GRACE to understanding climate change, Nat. Clim. Chang., 9(5), 358–369, doi:10.1038/s41558-019-0456-2, 2019.

Thomas, B. F., Famiglietti, J. S., Landerer, F. W., Wiese, D. N., Molotch, N. P. and Argus, D. F.: GRACE Groundwater Drought Index: Evaluation of California Central Valley groundwater drought, Remote Sens. Environ., 198, 384–392, doi:10.1016/j.rse.2017.06.026, 2017.

Thomas, T., Jaiswal, R. K., Nayak, P. C. and Ghosh, N. C.: Comprehensive evaluation of the changing drought characteristics in Bundelkhand region of Central India, Meteorol. Atmos. Phys., 127(2), 163–182, doi:10.1007/s00703-014-0361-1, 2015.

Tian, J., Li, C., Liu, J., Yu, F., Cheng, S., Zhao, N. and Wan Jaafar, W. Z.: Groundwater depth prediction using data-driven models with the assistance of gamma test, Sustain., 8(11), 1–17, doi:10.3390/su8111076, 2016.

Torrence, C. and Compo, G. P.: A Practical Guide to Wavelet Analysis, Bull. Am. Meteorol. Soc., 79(1), 61–78, doi:10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2, 1998.

Torrence, C. and Webster, P. J.: Interdecadal Changes in the ENSO–Monsoon System, J. Clim., 12(8), 2679–2690, doi:10.1175/1520-0442(1999)012<2679:ICITEM>2.0.CO;2, 1999.

Tyson, H. N. and Weber, E. M.: Ground-Water Management for the Nation's Future—Computer Simulation of Ground-Water Basins, J. Hydraul. Div., 90(4), 59–77, doi:10.1061/JYCEAJ.0001093, 1964.

Uddameri, V., Singaraju, S. and Hernandez, E. A.: Is Standardized Precipitation Index (SPI) a Useful Indicator to Forecast Groundwater Droughts? - Insights from a Karst Aquifer, JAWRA J. Am. Water Resour. Assoc., 55(1), 70–88, doi:10.1111/1752-1688.12698, 2019.

van der Gun, J.: Groundwater resources sustainability, in Global Groundwater, edited by A. Mukherjee, Abhijit; Scanlon, Bridget; Aureli, Alice; Langan, Simon; Guo, Huaming; McKenzie, pp. 331–345, Elsevier., 2020.

Van Lanen, H. A. J., Laaha, G., Kingston, D. G., Gauster, T., Ionita, M., Vidal, J., Vlnas, R., Tallaksen, L. M., Stahl, K., Hannaford, J., Delus, C., Fendekova, M., Mediero, L., Prudhomme, C., Rets, E., Romanowicz, R. J., Gailliez, S., Wong, W. K., Adler, M., Blauhut, V., Caillouet, L., Chelcea, S., Frolova, N., Gudmundsson, L., Hanel, M., Haslinger, K., Kireeva, M., Osuch, M., Sauquet, E., Stagge, J. H. and Van Loon, A. F.: Hydrology needed to manage droughts: the 2015 European case, Hydrol. Process., 30(17), 3097–3104, doi:10.1002/hyp.10838, 2016.

Van Loon, A. F., Kumar, R. and Mishra, V.: Testing the use of standardised indices and GRACE satellite data to estimate the European 2015 groundwater drought in near-real time, Hydrol. Earth Syst. Sci., 21(4), 1947–1971, doi:10.5194/hess-21-1947-2017, 2017.

Van Loon, A. F.: Hydrological drought explained, WIREs Water, 2(4), 359–392, doi:10.1002/wat2.1085, 2015.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I.: Attention Is All You Need, [online] Available from: http://arxiv.org/abs/1706.03762, 2017.

Veleda, D., Montagne, R. and Araujo, M.: Cross-wavelet bias corrected by normalizing scales, J. Atmos. Ocean. Technol., 29(9), 1401–1408, doi:10.1175/JTECH-D-11-00140.1, 2012.

Vicente-Serrano, S. M., Beguería, S. and López-Moreno, J. I.: A multiscalar drought index sensitive to global warming: The standardized precipitation evapotranspiration index, J. Clim., 23(7), 1696–1718, doi:10.1175/2009JCLI2909.1, 2010.

Vu, M. T., Jardani, A., Massei, N. and Fournier, M.: Reconstruction of missing groundwater level data by using Long Short-Term Memory (LSTM) deep neural network, J. Hydrol., 597, 125776, doi:10.1016/j.jhydrol.2020.125776, 2021.

Watkins, M. M., Wiese, D. N., Yuan, D.-N., Boening, C. and Landerer, F. W.: Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons, J. Geophys. Res. Solid Earth, 120(4), 2648–2671, doi:10.1002/2014JB011547, 2015.

Wilhite, D. A.: Drought as a natural hazard: Concepts and definitions, in Drought: A Global Assessment, vol. 1, Routledge, London, 3–18, 2000.

Wunsch, A., Liesch, T. and Broda, S.: Forecasting groundwater levels using nonlinear autoregressive networks with exogenous input (NARX), J. Hydrol., 567, 743–758, doi:10.1016/j.jhydrol.2018.01.045, 2018.

Wunsch, A., Liesch, T. and Broda, S.: Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX), Hydrol. Earth Syst. Sci., 25(3), 1671–1687, doi:10.5194/hess-25-1671-2021, 2021.

Yan, J., Mu, L., Wang, L., Ranjan, R. and Zomaya, A. Y.: Temporal Convolutional Networks for the Advance Prediction of ENSO, Sci. Rep., 10(1), 8055, doi:10.1038/s41598-020-65070-5, 2020.

Yang, C.-C., Prasher, S. O., Lacroix, R., Sreekanth, S., Patni, N. K. and Masse, L.: Artificial Neural Network Model for Subsurface-Drained Farmlands, J. Irrig. Drain. Eng., 123(4), 285–292, doi:10.1061/(ASCE)0733-9437(1997)123:4(285), 1997.

Yoon, H., Jun, S. C., Hyun, Y., Bae, G. O. and Lee, K. K.: A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer, J. Hydrol., 396(1–2), 128–138, doi:10.1016/j.jhydrol.2010.11.002, 2011.

York, J. P., Person, M., Gutowski, W. J. and Winter, T. C.: Putting aquifers into atmospheric simulation models: an example from the Mill Creek Watershed, northeastern Kansas, Adv. Water Resour., 25(2), 221–238, doi:10.1016/S0309-1708(01)00021-5, 2002.

ZAKHEM, B. A. and KATTAA, B.: Investigation of hydrological drought using Cumulative Standardized Precipitation Index (SPI 30) in the eastern Mediterranean region (Damascus, Syria), J. Earth Syst. Sci., 125(5), 969–984, doi:10.1007/s12040-016-0703-0, 2016.

Zhang, D., Lindholm, G. and Ratnaweera, H.: Use long short-term memory to enhance Internet of Things for combined sewer overflow monitoring, J. Hydrol., 556, 409–418, doi:10.1016/j.jhydrol.2017.11.018, 2018.

Zhang, J., Zhu, Y., Zhang, X., Ye, M. and Yang, J.: Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas, J. Hydrol., 561(January), 918–929, doi:10.1016/j.jhydrol.2018.04.065, 2018.

Zhang, X., Peng, Y., Zhang, C. and Wang, B.: Are hybrid models integrated with data preprocessing techniques suitable for monthly streamflow forecasting? Some experiment evidences, J. Hydrol., 530, 137–152, doi:10.1016/j.jhydrol.2015.09.047, 2015.

Zhao, T., Zhu, Y., Ye, M., Mao, W., Zhang, X., Yang, J. and Wu, J.: Machine-Learning Methods for Water Table Depth Prediction in Seasonal Freezing-Thawing Areas, Groundwater, 58(3), 419–431, doi:10.1111/gwat.12913, 2020.

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. Stefan Kollet, for the continuous guidance and support during my PhD study. When I encountered obstacles in my research, he always helped me with great patience and immense knowledge. The achievements of this PhD work would not have been possible without him. In addition, he offered me the opportunity to attend various conferences and seminars, which have helped me broaden my insights into hydrometeorological sciences.

Besides my supervisor, I would like to thank the rest of my doctoral committee, Prof. Clemens Simmer, Prof. Julian Klaus and Prof. Jürgen Kusche, for their engagement, insightful comments and good questions.

Also, I am grateful to my institute supervisor, Dr. Carsten Montzka, who reviewed Chapters 2, 3 and 4 and provided valuable comments for improving them. He gave me many supports in terms of administrative procedures, which significantly enhanced my working and studying experience at Forschungszentrum Jülich (FZJ).

I would like to thank all current and former members of my two research groups at FZJ for the fruitful collaboration, the pleasant working environment and the common activities unrelated to science. In particular, I am grateful to Dr. Bagher Bayat and Dr. Bibi Naz for their concrete suggestions on Chapters 2, 3 and 4. I would also like to thank Dr. Carina Furusho-Percot and Carl Hartick for providing the data to train my machine learning (ML) models and analyze results. Further, I am very thankful to Dr. Klaus Görgen for sharing his extensive knowledge of earth system modeling with me and to Tobias Tesch for many interesting discussions on ML.

Finally, I would like to thank my parents who have been very supportive and encouraging throughout my life. 爸爸妈妈，谢谢你们!

Band / Volume 572
Developing an integrated value-based institutional framework for analyzing nexus governance challenges – the case study of Germany
C. Märker (2022), 290 pp
ISBN: 978-3-95806-617-5

Band / Volume 573
Ecological sanitation via thermophilic co-composting of humanure and biochar as an approach to climate-smart agriculture
D. Castro Herrera (2022), XVIII, 127 pp
ISBN: 978-3-95806-622-9

Band / Volume 574
Towards 3D crosshole GPR full-waveform inversion
A. Mozaffari (2022), viii, 122 pp
ISBN: 978-3-95806-623-6

Band / Volume 575
Investigations of the atmospheric OH, HO2 and RO2 radical chemical budgets and their impact on tropospheric ozone formation in a rural area in West-Germany in the JULIAC 2019 campaign
C. Cho (2022), 182 pp
ISBN: 978-3-95806-625-0

Band / Volume 576
Thermochemische Beständigkeit von Carbonat-Keramik-Membranen für die CO_2-Abtrennung in Wassergas-Shift-Reaktoren
U. Gude (2022), X, 176, LXIII pp
ISBN: 978-3-95806-626-7

Band / Volume 577
Neue Ziele auf alten Wegen?
Strategien für eine treibhausgasneutrale Energieversorgung
bis zum Jahr 2045
D. Stolten, P. Markewitz, T. Schöb, F. Kullmann, S. Risch, T. Groß, M. Hoffmann, D. Franzmann, T. Triesch, S. Kraus, R. Maier, B. Gillessen, H. Heinrichs, N. Pflugradt, T. Grube, J. Linssen, L. Kotzur (2022), VI, 81 pp
ISBN: 978-3-95806-627-4

Band / Volume 578
Improving stationary and mobile cosmic ray neutron soil moisture measurements
Assessment of the cosmic ray neutron uncertainty
and the potential of the thermal neutron signal
J. C. Jakobi (2022), xxiii, 137 pp
ISBN: 978-3-95806-628-1

Band / Volume 579
Application-Specific Calibration of Condensation Particle Counters under Low Pressure Conditions
O. B. F. Bischof (2022), ix, 137 pp
ISBN: 978-3-95806-629-8

Band / Volume 580
Konzepte und Potenziale von Demonstrationsanlagen für die Produktion von erneuerbarem synthetischen Flugzeugtreibstoff als Beitrag zur Transformation der Reviere in NRW – Abschlussbericht
S. Weiske, F. Schorn, J. L. Breuer, L. Becka, N. Beltermann, N. Wegener, R. C. Samsun, P. Moser, S. Schmidt, C. Götte, E. Rietmann, U. Balfanz, A. Spieckermann, D. Beckmann, J. Edler-Krupp, F. Steffen und R. Peters (2022), 167 pp
ISBN: 978-3-95806-630-4

Band / Volume 581
Schlussbericht
Accelerating Low- carbon Industrial Growth through CCUS (ALIGN-CCUS)
S. Weiske, S. Schemme, R. C. Samsun, R. Peters, S. Troy, S. Morgenthaler, D. Schumann, P. Zapp (2022), 216 pp
ISBN: 978-3-95806-632-8

Band / Volume 582
Closed Carbon Cycle Mobility –
Klimaneutrale Kraftstoffe für den Verkehr der Zukunft
F. Schorn, J. Häusler, J. L. Breuer, S. Weiske, J. Pasel, R. C. Samsun, R. Peters (2022), 220 pp
ISBN: 978-3-95806-636-6

Band / Volume 583
Machine learning for monitoring groundwater resources over Europe
Y. Ma (2022), viii, 125 pp
ISBN: 978-3-95806-638-0

Weitere ***Schriften des Verlags im Forschungszentrum Jülich*** unter
http://wwwzb1.fz-juelich.de/verlagextern1/index.asp

Energie & Umwelt / Energy & Environment
Band / Volume 583
ISBN 978-3-95806-638-0

Mitglied der Helmholtz-Gemeinschaft

JÜLICH
Forschungszentrum