https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions

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

Yueling Ma[1,2], Carsten Montzka[1], Bagher Bayat[1], Stefan Kollet[1,2]

[1] Institute of Bio and Geosciences (Agrosphere, IBG-3), Research Center Jülich, 52428 Jülich, Germany
[2] Center for High-Performance Scientific Computing in Terrestrial Systems, Geoverbund ABC/J, 52428 Jülich, Germany

Correspondence to: Yueling Ma (y.ma@fz-juelich.de)

Abstract. Many European countries mainly rely on groundwater for domestic water use. Due to a scarcity of near real-time water table depth (*wtd*) observations, establishing a spatially consistent groundwater monitoring system at the continental scale is a challenge. Hence, it is necessary to develop alternative methods to estimate *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 to produce monthly wtd_a, using monthly precipitation anomalies (pr_a) as input. LSTM networks are a special category of artificial neural networks, useful in detecting a long-term dependency within sequences, in our case time series, which is expected in the relationship between pr_a and wtd_a. To set up the methodology, spatio-temporally continuous data were obtained from daily terrestrial simulations (hereafter termed the TSMP-G2A data set) with a spatial resolution of 0.11°, ranging from the year 1996 to 2016. They 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 modeled wtd_a maps from LSTM networks agreed well with TSMP-G2A wtd_a maps in 2003 and 2015 constituting drought years over Europe. Moreover, we categorized test performances of the networks based on yearly averaged *wtd*, evapotranspiration (*ET*), soil moisture (θ), snow water equivalent (S_w), and 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^{3}m^{-3} and S_w < 10 mm, revealing a significant impact of the local factors on the ability of the networks to process information. Furthermore, results of 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 to produce high-quality wtd_a based on other hydrometeorological data measured and predicted at large scales, such as *pr*. This contribution may facilitate the establishment of an effective groundwater monitoring system over Europe relevant to water management.

1 Introduction

Groundwater 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

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.

vulnerability to pollution (Naghibi et al., 2016; Tian et al., 2016). According to the report of the European Environment Agency (EEA) in 1999, groundwater comprises over 50% of public water supply in most European countries. Groundwater systems are dynamic and adapt continuously to natural and anthropogenic stresses (Kenda et al., 2018). However, they are affected in recent years as a consequence of frequent extreme weather conditions, e.g., severe droughts and human overexploitation. Thus, effective and efficient groundwater 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, and more droughts have occurred in some European countries than in the past, and their severity has also been increased. Recent severe heatwave events in Europe occurred in 2003, 2015, and 2018, which lead to several drought events covering most of the European continent (Norris, 2018). Groundwater 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, groundwater monitoring is ultimately indispensable over the European continent.

Effective groundwater monitoring requires accurate information on groundwater dynamics in space and time. One crucial variable for characterizing groundwater dynamics is water table depth anomaly (wtd_a), reflecting anomalies in groundwater storage (Zhao et al., 2020), which is a key variable in groundwater drought analysis. wtd_a is derived from *wtd* observations, which are measured directly from observation wells. However, to date, there 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 for producing reliable area-wide wtd_a information over Europe.

Indirect methods rely on measurements of one or more hydrometeorological variables related to *wtd* via physical processes in the water cycle, such as infiltration and percolation. Information regarding precipitation anomaly (pr_a) is the most common variable used to model wtd_a. Precipitation (*pr)* is mostly connected with groundwater 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 groundwater to *pr*. A considerable number of studies linked the accumulation of pr_a over extended time scales (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 weights were 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, having many interconnected information-processing units (i.e., neurons) (Haykin, 2009; Ma et al., 2019). By adapting learnable parameters (i.e., weights and biases) on the links between neurons, ANNs can give an appropriate input-output mapping based on observed data even for complex

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.
Hydrology and Earth System Sciences
Open Access
EGU
Discussions

nonlinear relationships. ANNs are not easily affected by input noise and able to readjust their parameters when new information is included. More importantly, compared to physically-based models, they necessitate little background knowledge, reducing the requirements for human involvement and expertise, and thus, enabling rapid hypothesis testing (Govindaraju, 2000; Shen, 2018; Sun and Scanlon, 2019).

Recurrent neural networks (RNNs) are 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 more efficiently solve sequential data problems such as groundwater level estimation than feedforward networks and their variants. The latter are commonly used ANNs for groundwater level modeling in previous studies, e.g., Yang et al. (1997), Nayak et al. (2006), Adamowski and Chan (2011), Yoon et al. (2011), Gong et al. (2015), Mohanty et al. (2015), Sun et al. (2016). With RNNs, it is not necessary to estimate the delay time d (d > 0) in the network response in advance and to assign one input variable to several input neurons (namely, the input data at the time steps t, t – 1, …, t - d + 1) during modeling like it is with feedforward networks (J. Zhang et al., 2018; Supreetha et al., 2020).

Long Short-Term Memory (LSTM) networks are a special type of RNNs and famous because of their superior performance in exploiting long-term dependencies between sequences, which is expected in the response of *wtd* to *pr*. Although LSTM networks have been employed extensively in other science fields, particularly 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 groundwater fluctuations using 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 to visualize 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) and Banerjee and Mitra (2014).

In this study, we utilized spatio-temporally continuous pr_a and wtd_a from integrated hydrologic simulation results over Europe (hereafter termed TSMP-G2A data set, introduced in Sect. 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 (S_w), 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 paper is organized as follows: in Sect. 2 (Methodology), we first present a conceptual model of groundwater balance to theoretically derive the relationship between pr_a and wtd_a and then briefly introduce the architecture of the proposed LSTM networks, continuous and cross-wavelet transform. This is then followed by detailed information of our study area

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.

and data set as well as a generic workflow to construct local LSTM networks at selected pixels over Europe. Section 3 (Results and discussion) shows reproduced wtd_a maps for groundwater drought analysis, discusses the impact of local factors on the network behaviors and investigates the network performances at the local scale, before completing the paper with Sect. 4 (Summary and conclusions).

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 groundwater balance, introduce the principle of LSTM networks and the application of XWT, and describe the study area and data set and a universal workflow to establish the proposed LSTM networks locally at selected pixels.

2.1 Conceptual model of groundwater balance

The complete subsurface water balance can be described by a control volume that contains the vadose zone, and an unconfined aquifer closed at the bottom (Fig. 1). Areas with surface water are not taken into account. Fluxes in and out of the control volume are *pr* and *ET* across the land surface and lateral fluxes in the subsurface. These fluxes are balanced by changes in the water stored in the vadose zone and the unconfined aquifer.

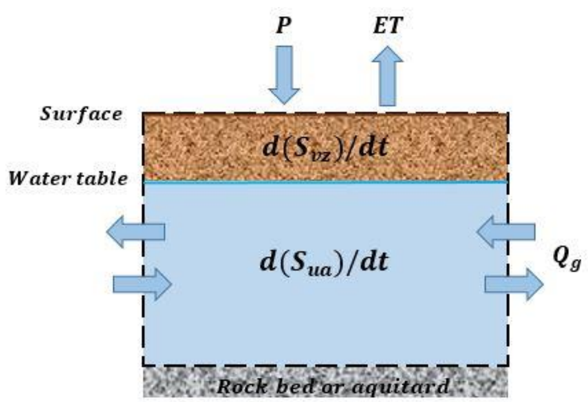

Figure 1: Conceptual model of groundwater balance over a control volume. Variables are defined in the text (modified from Maxwell, 2010).

The complete groundwater balance equation for the conceptual model is given in Eq. (1):

$$d(S_{vz})/dt + d(S_{ua})/dt = P - ET + div(\boldsymbol{Q_g}) \, . \tag{1}$$

Rearranging Eq. (1), will result in Eq. (2) as follows:

$$d(S_{ua})/dt = P - ET + div\left(\boldsymbol{Q_g}\right) - d(S_{vz})/dt \, , \tag{2}$$

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.
Hydrology and Earth System Sciences Discussions

where, P is precipitation [LT^{-1}]; and ET is actual evapotranspiration [LT^{-1}]; and $div(\boldsymbol{Q_g})$ is the divergence of groundwater [LT^{-1}]; and S_{vz} and S_{ua} are the water storages in the vadose zone [L] and the unconfined aquifer [L], respectively; and t is time [T].

The term on the left-hand side and the first term on the right-hand side in Eq. (2) indicate an explicit relationship between the fluctuation of S_{ua} and *pr*, providing the theoretical basis of this study. In case of large continental watersheds (i.e., $div(\boldsymbol{Q_g}) = 0$), the difference between *pr* and *ET* is equal to the total variations in S_{vz} and S_{ua}. Note, 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 (*P*-*ET*).

2.2 Long Short-Term Memory networks

In this study, we employed LSTM networks having the same architecture of hidden neurons as Gers et al. (2000). As a category of RNNs, LSTM networks have loops in their hidden layers, facilitating hidden neurons to weigh not only new inputs but also earlier outputs internally for predictions. Hence, 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 Fig. 2), in order to overcome the gradient exploring and vanishing issue. 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 in 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 procedure for processing inputs in hidden neurons of LSTM networks are 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 remember 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; 4) compute outputs of hidden neurons from the new CEC state and the results given by the output gate.

Figure 2 illustrates a one-hidden-layer LSTM network containing only one hidden neuron; the pseudocode is presented in Appendix A. Owing to limited data available at each pixel (i.e., a total of 252 time steps), we built small LSTM networks at the local scale, having 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. 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 of

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.

processing new data, and it can be a double-edged sword to apply a network with considerable hidden neurons. On the one hand, the larger 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, 2017). 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 hyperparameter tuning described in Sect. 2.5.

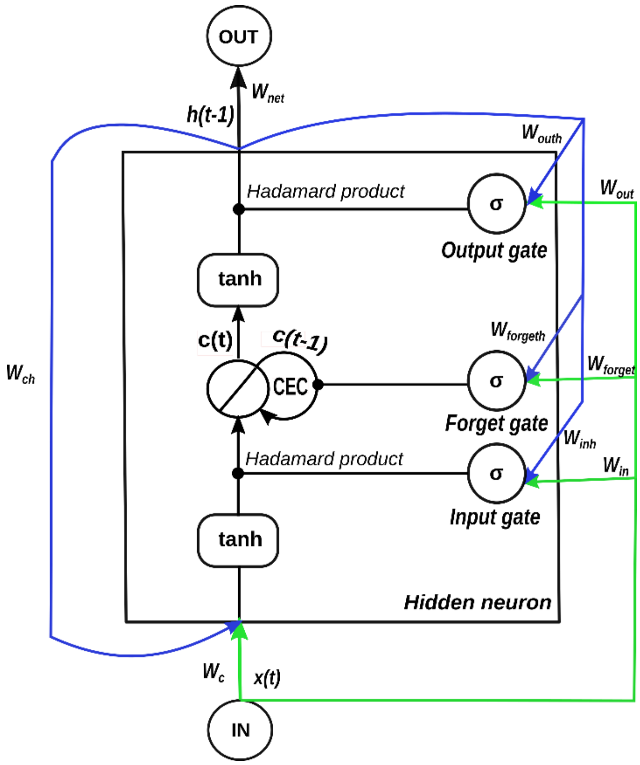

Figure 2: One-hidden-layer LSTM network with one hidden neuron. The green lines indicate the entry points of new inputs into the hidden neuron. The blue lines show the entry points of previous outputs into the hidden neuron, where w_* is the weight on a linkage; $h(*)$ is the output of the hidden neuron; $x(t)$ is the input at the time step t; and $c(*)$ is the cell state. σ represents a sigmoid function, and tanh is a hyperbolic tangent function.

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 $\psi_0(\eta)$, η being a dimensionless time parameter, the CWT of a time series x_{n0} is formulated as the convolution of x_{n0} and a scaled and translated form of $\psi_0(\eta)$ (Torrence and Compo, 1998):

$$W(s,n) = \sum_{n0=0}^{N-1} x_{n0}\psi^*[(n0-n)\delta t/s] \ , \quad (3)$$

where, the (*) signifies the complex conjugate; δt is the time step of x_{n0}; N is the total number of δt in x_{n0}; 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 $|W(s,n)|^2$.

Hydrology and Earth System Sciences
Open Access
Discussions

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:

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

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

XWT is a method to locate common high power in the wavelet transforms of two time series. The XWT of two time series x_{n0} and y_{n0} can be computed using (Grinsted et al., 2004):

$$W_{xy}(s,n) = W_x(s,n) W_y^*(s,n) , \quad (5)$$

where, $W_x(s,n)$ and $W_y(s,n)$ are the CWT of time series x_{n0} and y_{n0}, 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 here 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, the application of XWT aims to identify common, localized high-power frequency modes of $\psi_0(\eta)$ in input and expected output series and detect dynamics of the modes over time. Using the XWT analysis, we expect to clarify whether a changing pattern exists in the input-output 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 internal operations of LSTM networks.

2.4 Study area and data set

We constructed the LSTM networks at individual pixels over eight hydrometeorologically different regions within Europe (Fig. 3), which are known as the PRUDENCE regions (Christensen and Christensen, 2007). Table 1 lists 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 S_w of which data are only available from the years 2003 to 2010. The TSMP-G2A data set consists of daily averaged simulation results from the Terrestrial Systems Modeling Platform (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 groundwater via the land surface to the top of the atmosphere (Keune et al., 2016; Furusho-Percot et al., 2019). 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) and Keune et al. (2019). Furusho-Percot et al. (2019) compared simulated anomaly data of temperature, *pr*, and total column water storage from TSMP with commonly used reference datasets (i.e., the 0.25 degrees gridded European Climate Assessment

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.
Hydrology and Earth System Sciences
Open Access
Discussions

and Dataset, E-OBS v19, ECA&D, and observations from the Gravity Recovery and Climate Experiment, GRACE), showing good agreement between the simulated and observed values. For details of the TSMP-G2A data set, the reader is referred to Furusho-Percot et al. (2019).

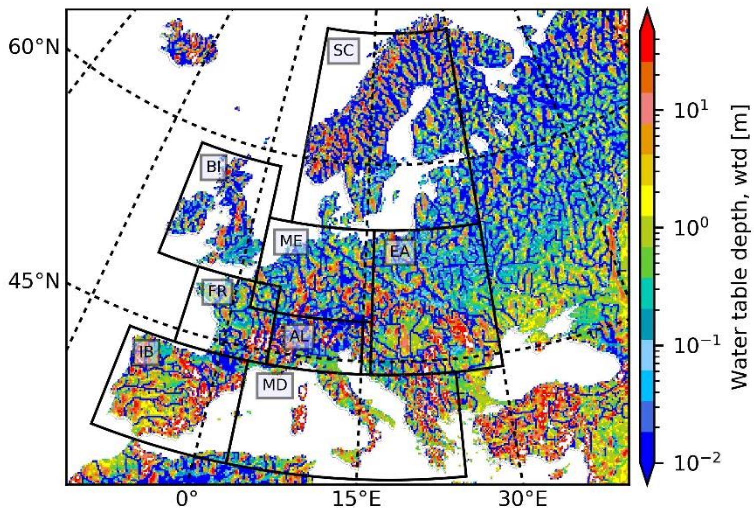

Figure 3: TSMP-G2A *wtd* [m] climatology over the European continent for the time period 1996 to 2016. Areas bounded by the thick black lines show the PRUDENCE regions (i.e., SC: Scandinavia; BI: British Isles; ME: Mid-Europe; EA: Eastern Europe; FR: France; AL: Alps; IB: Iberian Peninsula; MD: Mediterranean).

As shown by the averages in Table 1, *pr* is heterogeneously distributed over the PRUDENCE regions, with the highest rainfall in AL (1480 mm) and the lowest in EA (778 mm). Most regional average *wtd* ranges from 2 m to 4 m, other than IB and MD (having a larger average *wtd* > 6 m). Within this range, AL has a relatively high average *wtd* (3.95 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 IB (0.28 m^3m^{-3}) and MD (0.29 m^3m^{-3}). For S_w, large values (> 60 mm) are simulated in SC and AL, while values below 10 mm are recorded in the other regions.

Table 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 (lon_west, lon_east, lat_south, lat_north)	Regional precipitation, *pr* [mm]	Regional water table depth, *wtd* [m]	Regional evapotranspiration, *ET* [mm]	Regional soil moisture, θ [m^3m^{-3}]	Regional snow water equivalent, S_w [mm]
(SC) Scandinavia	(5, 30, 55, 70)	1007 ± 454	2.32 ± 5.56	283 ± 129	0.31 ± 0.11	79.80 ± 109.17
(BI) British Isles	(-10, 2, 50, 59)	1120 ± 308	2.18 ± 5.83	395 ± 130	0.34 ± 0.10	0.82 ± 2.19

© Author(s) 2020. CC BY 4.0 License.

(ME) Mid-Europe	(2, 16, 48, 55)	896 ± 211	2.64 ± 6.55	444 ± 141	0.33 ± 0.10	2.44 ± 5.49
(EA) Eastern Europe	(16, 30, 44, 55)	778 ± 187	2.94 ± 7.03	470 ± 164	0.32 ± 0.10	9.50 ± 13.07
(FR) France	(-5, 5, 44, 50)	895 ± 171	2.82 ± 6.72	485 ± 164	0.33 ± 0.09	0.31 ± 1.12
(AL) Alps	(5, 15, 44, 48)	1480 ± 644	3.95 ± 8.75	499 ± 185	0.34 ± 0.09	65.57 ± 127.23
(IB) Iberian Peninsula	(-10, 3, 36, 44)	841 ± 372	6.32 ± 10.13	495 ± 233	0.28 ± 0.11	3.38 ± 28.18
(MD) Mediterranean	(3, 25, 36, 44)	896 ± 340	6.19 ± 10.27	518 ± 229	0.29 ± 0.10	3.59 ± 15.22

We utilized the TSMP-G2A data set to compute pr_a and wtd_a (Eqs. (6)-(7)), 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 2.5) to guarantee that no future information leaks into the networks in the training process.

$$pr_a = (pr_m - pr_{av})/pr_{sd} \,, \tag{6}$$

where, pr_m is monthly sum pr 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); pr_{sd} is the climatological standard deviation of pr_m.

$$wtd_a = (wtd_m - wtd_{av})/wtd_{sd} \,, \tag{7}$$

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

To identify the effect of local factors on the network behaviors, we categorized the network performances based on yearly averaged *wtd*, *ET*, θ, S_w, 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 S_w data only available from 2003 to 2010), to ensure that spinup effects do not impact the analyses (Furusho-Percot et al., 2019).

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* to *pr*.

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

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.

al., 2017), stream discharge modeling (Zhang et al., 2015) and groundwater 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 has been introduced in Sect. 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 constant (Table 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: 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, learning rate	RMSprop (Hinton et al., n.d.), 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 individual deficiencies of the selected networks, thereby improving the quality of the final results (Goodfellow et al., 2017; Brownlee, 2018). The metrics to assess network performance in this study are the root mean square error (RMSE), the coefficient of determination (R^2) and the bias from the Pearson product-moment correlation coefficient R (α) as shown in Eqs. (8)-(10), respectively. α indicates systematic additive and multiplicative biases in the generated values, having a value between 0 and 1, where α = 1 means no bias (Duveiller et al., 2016).

$$RMSE = \sqrt{\sum_{i=1}^{N}\left(y_{exp} - y_{gene}\right)^2 / N}\,, \tag{8}$$

$$R^2 = 1 - \sum_{i=1}^{N}\left(y_{exp} - y_{gene}\right)^2 / \sum_{i=1}^{N}\left(y_{exp} - \overline{y_{exp}}\right)^2, \tag{9}$$

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.
Hydrology and Earth System Sciences
Open Access
EGU
Discussions

$$\alpha = 2 \,/\, \left[\sigma_{y_{exp}}/\sigma_{y_{gene}} + \sigma_{y_{gene}}/\sigma_{y_{exp}} + \left(\overline{y_{exp}} - \overline{y_{gene}}\right)^2 / \left(\sigma_{y_{exp}}\sigma_{y_{gene}}\right)\right], \quad (10)$$

where, y_{exp}, $\overline{y_{exp}}$, $\sigma_{y_{exp}}$ are the expected value, the average of the expected values, and the standard deviation of the expected values, respectively; y_{gene}, $\overline{y_{gene}}$, $\sigma_{y_{gene}}$ are the generated value, the average of the generated values, and the standard deviation of the generated 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 Fig. 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 Sect. 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 various groups based on yearly averaged *wtd*, *ET*, θ, S_w, and S_t and dominant *PFT* (Table 3), and the analysis result will be presented in Sect. 3.2. Figure 4 gives a generic workflow of this study to establish the LSTM networks at the local scale and analyze their output.

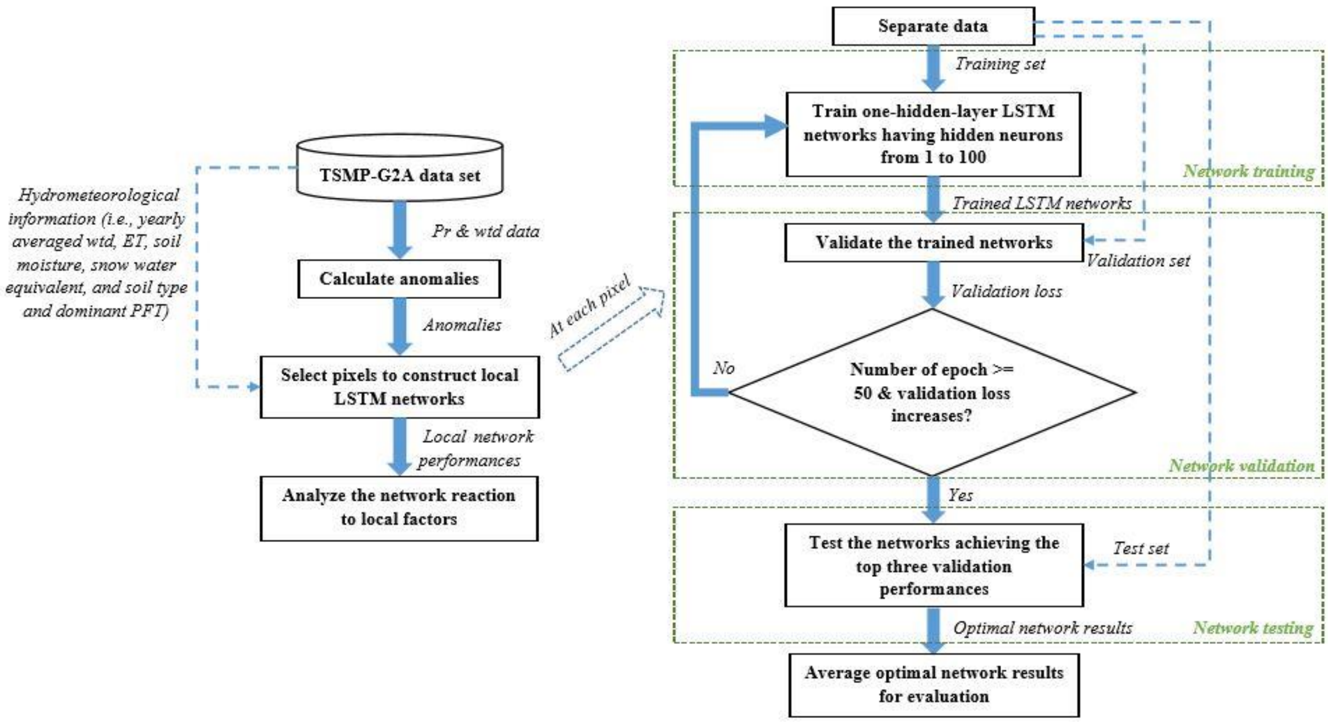

Figure 4: Workflow for LSTM network setup over the European CORDEX domain. The left section represents the overall processes of the network setup, whereas the right section shows how to apply LSTM networks at a selected pixel. The dash lines with arrows indicate additional data transmission paths.

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.

Table 3: Value ranges of yearly averaged *wtd, ET, θ,* S_w, 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^3m^{-3}]	Yearly averaged snow water equivalent, S_w [mm]	Soil type, S_t	Dominant plant functional type, *PFT**
1) 0.0–1.0; 2) 1.0-2.0; 3) 2.0-3.0; 4) 3.0-4.0; 5) 4.0-5.0; 6) 5.0-6.0; 7) 6.0-7.0; 8) 7.0-8.0; 9) 8.0-9.0; 10) 9.0-10.0; 11) 10.0-50.0.	1) < 0.0; 2) 0.0-100.0; 3) 100.0-200.0; 4) 200.0-300.0; 5) 300.0-400.0; 6) 400.0-500.0; 7) 500.0-600.0; 8) 600.0-700.0; 9) 700.0-800.0; 10) 800.0-900.0; 11) 900.0-1000.0; 12) 1000.0-1100.0.	1) 0.0-0.05; 2) 0.05-0.10; 3) 0.10-0.15; 4) 0.15-0.20; 5) 0.20-0.25; 6) 0.25-0.30; 7) 0.30-0.35; 8) 0.35-0.40; 9) 0.40-0.45; 10) 0.45-0.50.	1) ≤ 10.0 2) > 10.0	1) Sand; 2) loamy sand; 3) sandy loam; 4) silt loam; 5) silt; 6) loam; 7) sandy clay loam; 8) silty clay loam; 9) clay loam; 10) sandy clay; 11) silty clay; 12) clay; 13) organic material; 14) water; 15) bedrock; 16) others.	1) Needleleaf evergreen temperate tree; 2) needleleaf evergreen boreal tree; 3) needleleaf deciduous boreal tree; 4) broadleaf evergreen tropical tree; 5) broadleaf evergreen temperate tree; 6) broadleaf deciduous tropical tree; 7) broadleaf deciduous temperate tree; 8) broadleaf deciduous boreal tree; 9) broadleaf evergreen shrub; 10) broadleaf deciduous temperate shrub; 11) broadleaf deciduous boreal shrub; 12) c3 arctic grass; 13) c3 non-arctic grass; 14) c4 grass; 15) corn; 16) wheat.

**Dominant PFT: the PFT of which percentage is ≥ 50% at a pixel.*

3 Results and discussion

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). Figure 5 presents reproduced wtd_a maps over Europe for August 2003 and August 2015. The year 2003 is included in the training period, and as a result, the drought map derived from the wtd_a modeled from the networks (hereafter called LSTM wtd_a) is almost identical to the one based on the TSMP-G2A wtd_a (Fig. 5a). Apparently, a groundwater drought (i.e., $wtd_a \geq 1.5$) covered large parts of Europe, which is in good agreement with previous studies (Andersen et al., 2005; Van Loon et al., 2017). In the simulation, over central Germany, central Britain, southeastern France, the west Iberian Peninsula, and several parts in Eastern Europe, groundwater storage increased, illustrating the strong spatial heterogeneity of the anomalies, which is expected. In contrast, the year 2015 is part of the testing period, leading to a reduced agreement between the LSTM and TSMP-G2A wtd_a (Fig. 5b). Especially extremes in wet and dry anomalies were underestimated suggesting that the training set contains too little information on extreme events and, thus, is too short. Yet overall, visual inspection of Fig. 5b shows that the LSTM anomalies agree well with the expected values spatially, 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 Appendix B, leading to similar conclusions regarding the ability of the LSTM results to reproduce TSMP-G2A wtd_a.

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.
Hydrology and Earth System Sciences
Open Access
Discussions

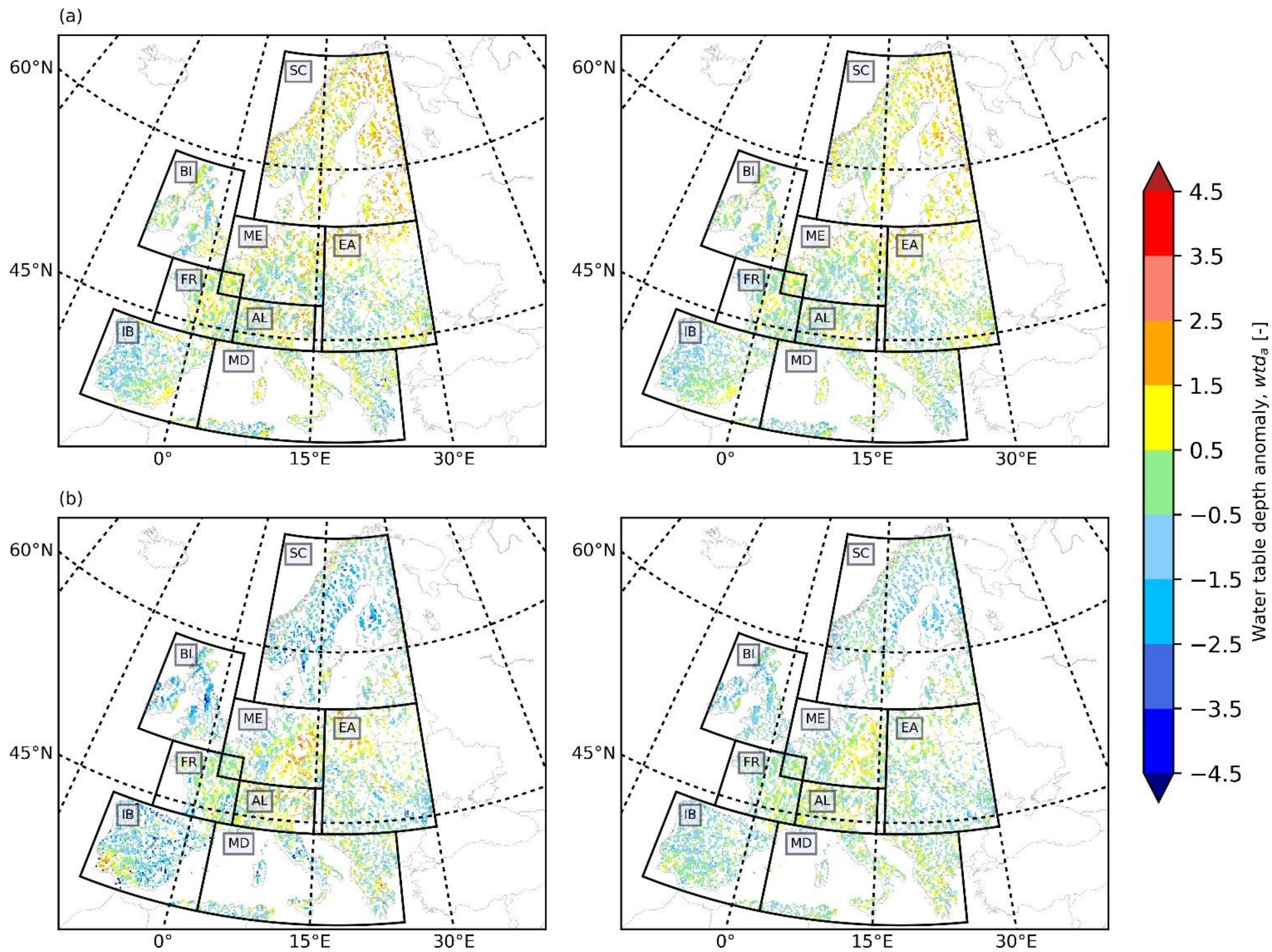

Figure 5: European wtd_a maps for (a) August 2003; (b) August 2015, derived from the TSMP-G2A data set (left) and results from LSTM networks (right).

3.2 Impact of local factors on the network performance

310 In each PRUDENCE region, we computed averages, and standard deviations of the test R^2 scores and RMSEs for the different categories (Table 3) of yearly averaged *wtd*, *ET*, θ, S_w, and S_t and dominant *PFT* (Fig. 6), to study dependents of the network test performances on different local factors. Note that negative values were set to zero in this calculation.

© Author(s) 2020. CC BY 4.0 License.

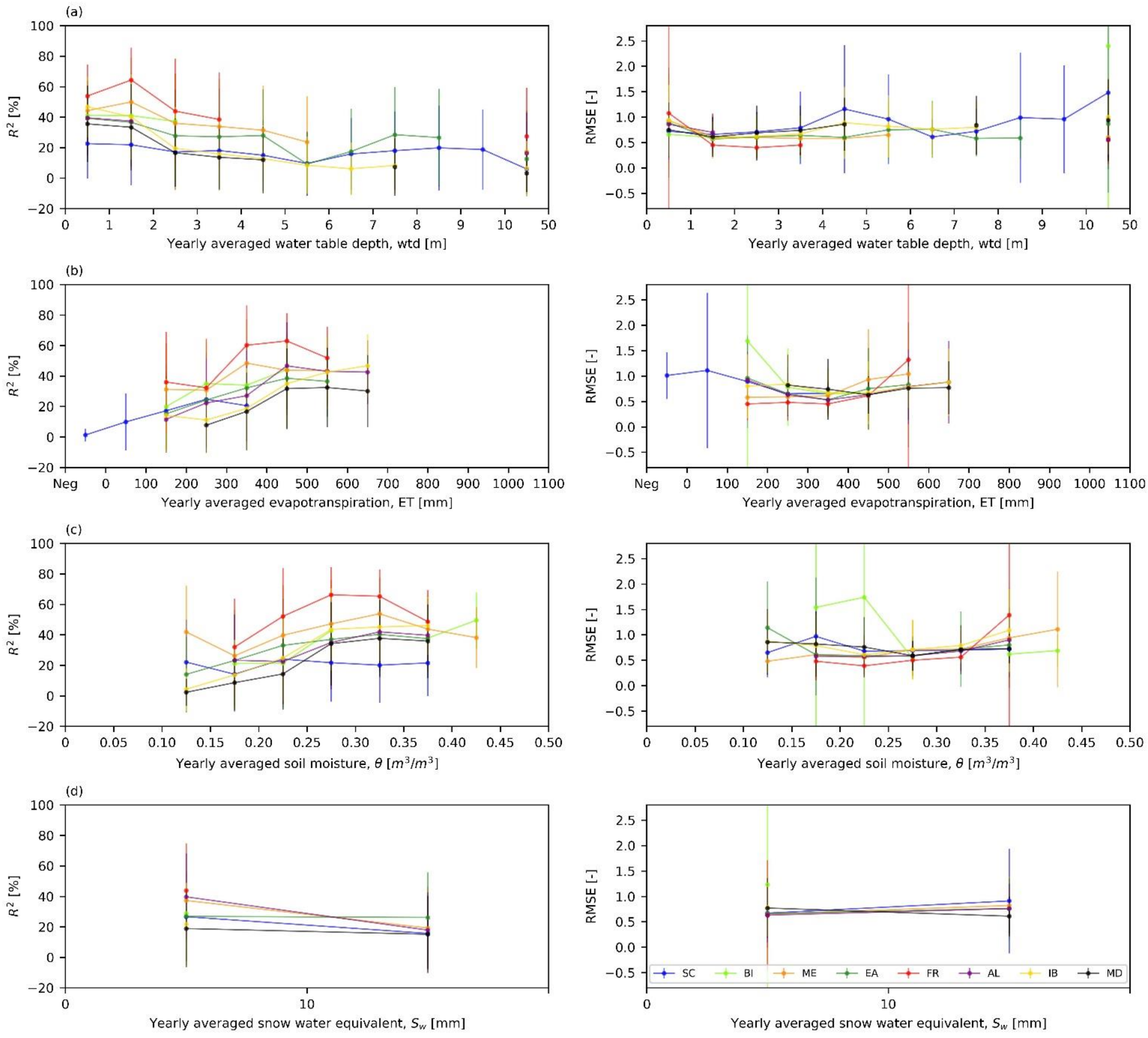

Figure 6: Averages and standard deviations of the test R^2 scores (left) and RMSEs (right) over the categorized results: yearly averaged (a) *wtd*; (b) *ET*; (c) θ; (d) S_w. The averages are indicated as dots, while the bars indicate standard deviations. The different colors reflect test results in different PRUDENCE regions.

There was no significant influence of S_t and dominant *PFT* on the scores. In general, the performance decreased with increasing yearly averaged *wtd*, which was manifested by decreasing average R^2 scores and growing average RMSEs (Fig. 6a). This type of network behavior can be attributed to a stronger connection of groundwater to *pr* 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* (Fig. 6b) or θ (Fig. 6c), there was an increase of

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.

average R^2 scores and a decrease of 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 S_w played an important role in the network test performance. In most PRUDENCE regions, the performance decreased in the case of S_w, leading to smaller average R^2 scores and larger average RMSEs presented in Fig. 6d. Snow accumulation resulted in complex feedback with groundwater processes that cannot be captured well by the networks without including additional input information.

As mentioned in Sect. 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, suggesting that a varying pattern may exist between pr_a and wtd_a over the study period (see Sect. 3.3).

Due to differences in the hydrometeorological characteristics of the PRUDENCE regions (see Table 1), the strength of the relationship between pr_a and wtd_a differed regionally (Fig. 6), leading to various regional test performances of the proposed LSTM networks shown in Fig. 7. Table 4 provides percentages of the selected pixels with test $R^2 \geq 50\%$ in each region, where FR exhibits the overall best network performance. Reduced performance in other regions appeared to be mainly related to high S_w in SC and AL and generally large *wtd* in IB and MD.

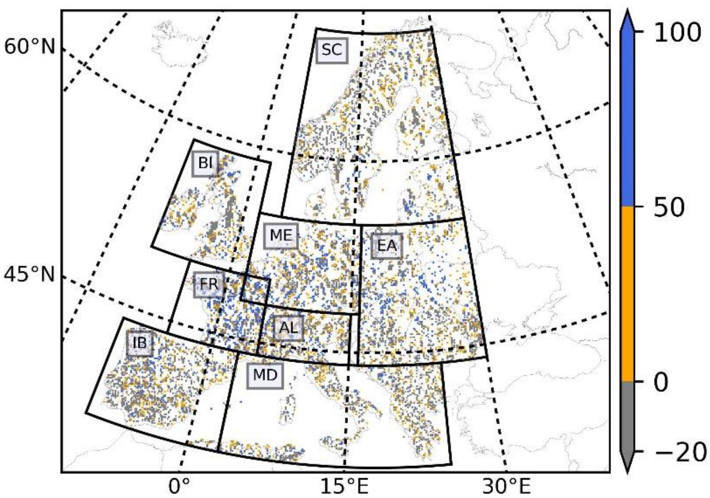

Figure 7: Map of test R^2 scores achieved by the proposed LSTM networks in the PRUDENCE regions.

Table 4: Percentages of the studied pixels with a test R^2 score ≥ 50% in the PRUDENCE regions [%].

SC	BI	ME	EA	FR	AL	IB	MD
14.08	30.77	39.30	27.27	52.84	30.30	22.46	16.33

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.

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

- C1: training R^2 score ≥ 50%, test R^2 score ≥50%;
- C2: training R^2 score ≥ 50%, test R^2 score ≤ 0%;
- C3: training R^2 score ≤ 0%, test R^2 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 and C3). C1 is the network behavior with satisfactory training and test scores. In contrast, in the case of C3, training and testing scores were unsatisfactory. An exception is C2, in which the networks that performed well on the training set failed to perform during testing. 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.

The distribution of the performance combinations was influenced by local factors. Figure 8 illustrates percentages of the pixels where the network performance followed C1 to C3 categorized by region, yearly averaged *wtd*, *ET*, θ *and* S_w, respectively. Each combination was required to have ≥ 50 study pixels. C1 was mainly found in areas with shallow *wtd*, high *ET*, high θ and little S_w, and was also the most common network performance in FR. In contrast, C2 mostly appeared in areas with large *wtd*, small *ET* and heavy S_w. This was a typical network performance at pixels with negative *ET* in SC, where processes such as freezing and sublimation were more pronounced than others due to low temperature. In addition, C3 appeared in a few pixels with *wtd* ≥ 10 m.

© Author(s) 2020. CC BY 4.0 License.

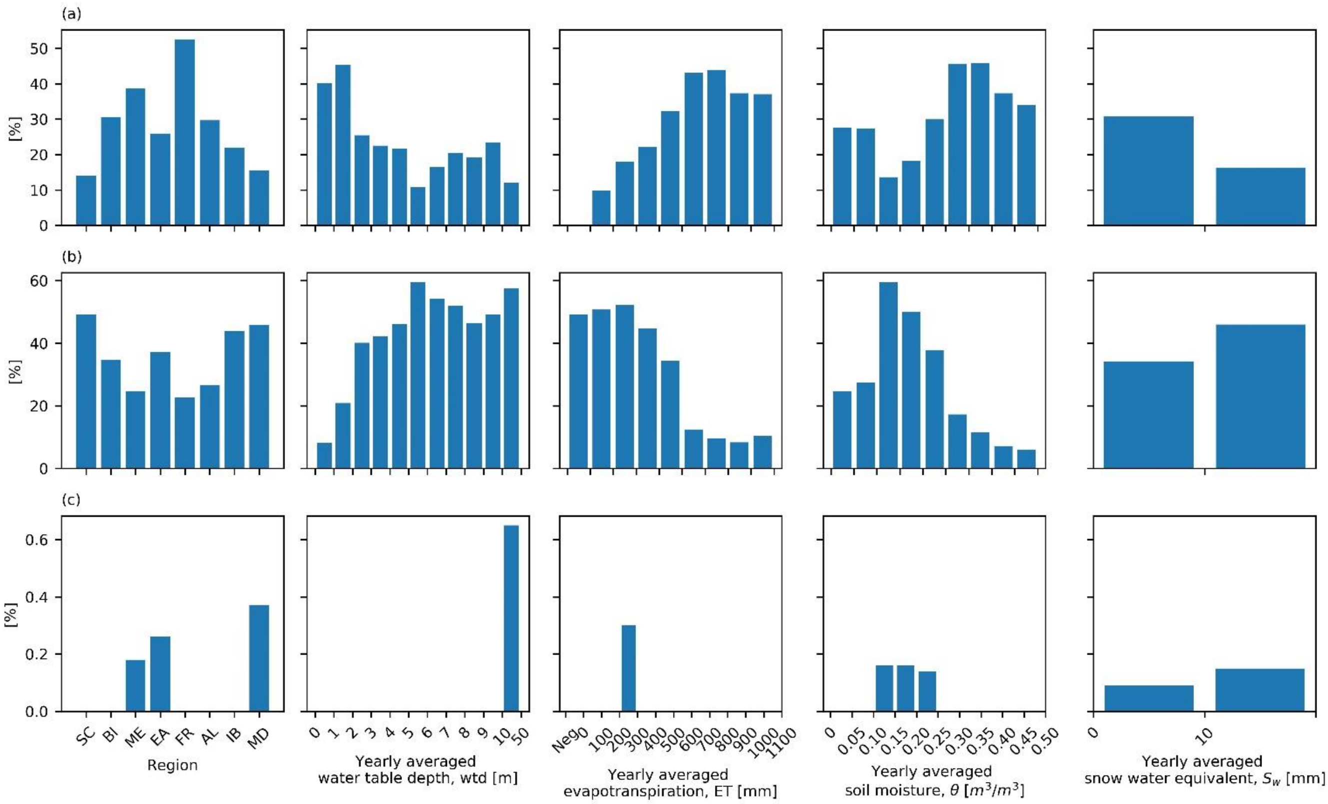

Figure 8: Histograms showing percentages of pixels where the network performance followed the combinations (a) C1; (b) C2; (c) C3. The plots show the combinations categorized by region, yearly averaged *wtd, ET, θ,* S_w, from left to right, respectively.

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 individual, representative pixels (Pixels 1-3, Table 5). XWT extracted the time localized coherence of the variability 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. The α values (Eq. (10)) of Pixel 1 were generally suggesting that smaller biases existed in the results of the LSTM networks. In addition, we found very different α values for Pixel 2 with small biases in the training and large biases in the validation and testing. The α values of Pixel 3 were generally small, indicating that the biases were large in the case of C3.

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.

Table 5: Pixel characteristics in the XWT analysis (Pixels 1-3).

	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, S_w [mm]
Pixel 1	C1	FR	1.32	422.91	0.27	0.0
Pixel 2	C2	SC	4.95	-24.41	0.15	535.0
Pixel 3	C3	MD	33.20	208.62	0.14	3.0

	Training R^2 [%]	Training α [%]	Validation R^2 [%]	Validation α [%]	Test R^2 [%]	Test α [%]
Pixel 1	73.20	70.71	66.43	47.07	74.94	70.08
Pixel 2	61.64	70.55	-66.38	7.27	-485.04	21.45
Pixel 3	-0.23	1.75	-30.63	3.89	-260.14	2.28

Figure 9 shows the results of the XWT analyses of the selected pixels in combination with the corresponding TSMP-G2A *pr* and *wtd* time series. Inspecting the results of the XWT analyses (bottom panel of Fig. 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 major defect of such data-driven models, so physically-based models cannot be completely replaced by data-driven models in this sense.

The power in the XWT results at Pixel 1 and Pixel 3 (Figs. 9a and 9c) was both nearly consistently located in a certain period, indicating an analogous pattern between pr_a and wtd_a throughout the whole study period, which is the prerequisite of good network performances. However, the networks behaved differently at the two pixels. By linking the XWT results to the associated network performances, we found that the networks tended to perform well when most of the power in the XWT results was consistently concentrated in the period from 2 to 16 months during the study period (see Fig. 9a). Supplementary plots in Appendix C showed similar phenomena as above. Therefore, we speculate that LSTM networks might be frequency-aware and work well on high-frequency components.

© Author(s) 2020. CC BY 4.0 License.

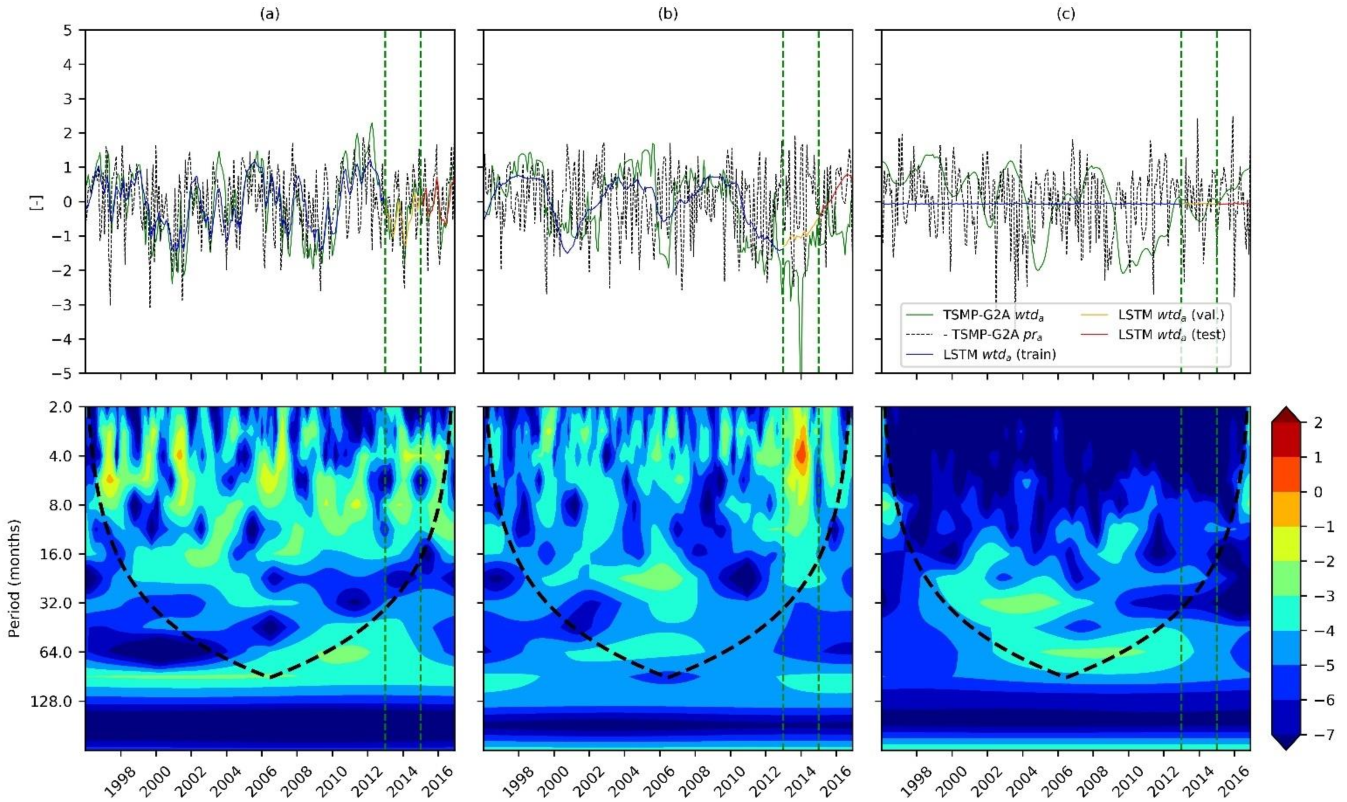

Figure 9: -TSMP-G2A *pr*pr_a, TSMP-G2A wtd_a and LSTM wtd_a time series (top) as well as cross-wavelet spectra for TSMP-G2A pr_a and wtd_a series (bottom) at a representative pixel of the performance combination (a) C1; (b) C2; (c) C3. In the cross-wavelet spectra, the black dashed line marks the boundary of the cone of influence; the color bar presents $log_2(power/scale)$. In all plots, the two green dashed lines separate the study period into the training, validation and testing periods.

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 1996 to 2018 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). 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, manifested by R^2 scores and RMSEs. Most of the networks attained high test R^2 scores at the pixels with wtd < 3 m, ET > 200 mm, θ > 0.15 m^3m^{-3} and S_w < 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 gave a hypothesis that LSTM networks have frequency awareness and tend to perform well on high-frequency components.

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. The results suggest that LSTM networks are useful to estimate wtd_a time series based on pr_a, which are routinely measured and, therefore, are more easily available from e.g., atmospheric re-analyses and forecast data sets and observations than groundwater level measurements. The proposed methodology may be transferred into a real-time monitoring and forecasting workflow for wtd_a at the continental scale.

Code and data availability. The code for constructing the proposed LSTM networks and result analyses is available from the authors. Please contract Yueling Ma at y.ma@fz-juelich.de. The TSMP-G2A data set is available online at https://doi.org/10.17616/R31NJMH3 (Furusho-Percot et al., 2019).

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.

Appendix A: Pseudocode of the LSTM network displayed in Fig. 2

Hereafter gives pseudocode of the one-hidden-layer LSTM networks illustrated in Fig. 2, which is modified from Gers et al. (2000). Variables were defined in the caption of Fig. 2. Note that, to simplify the code, biases are not shown here.

```
RESET all network parameters (i.e., weights, biases and cell states) as listed in Table 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) = σ(net_in(t))
        forget gate: f(t) = σ(net_forget(t))
        output gate: o(t) = σ(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., n.d.)
UNTIL validation error begins to drop and number of epochs ≥ 50
```

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.

Appendix B: Supplementary European water table depth anomaly maps

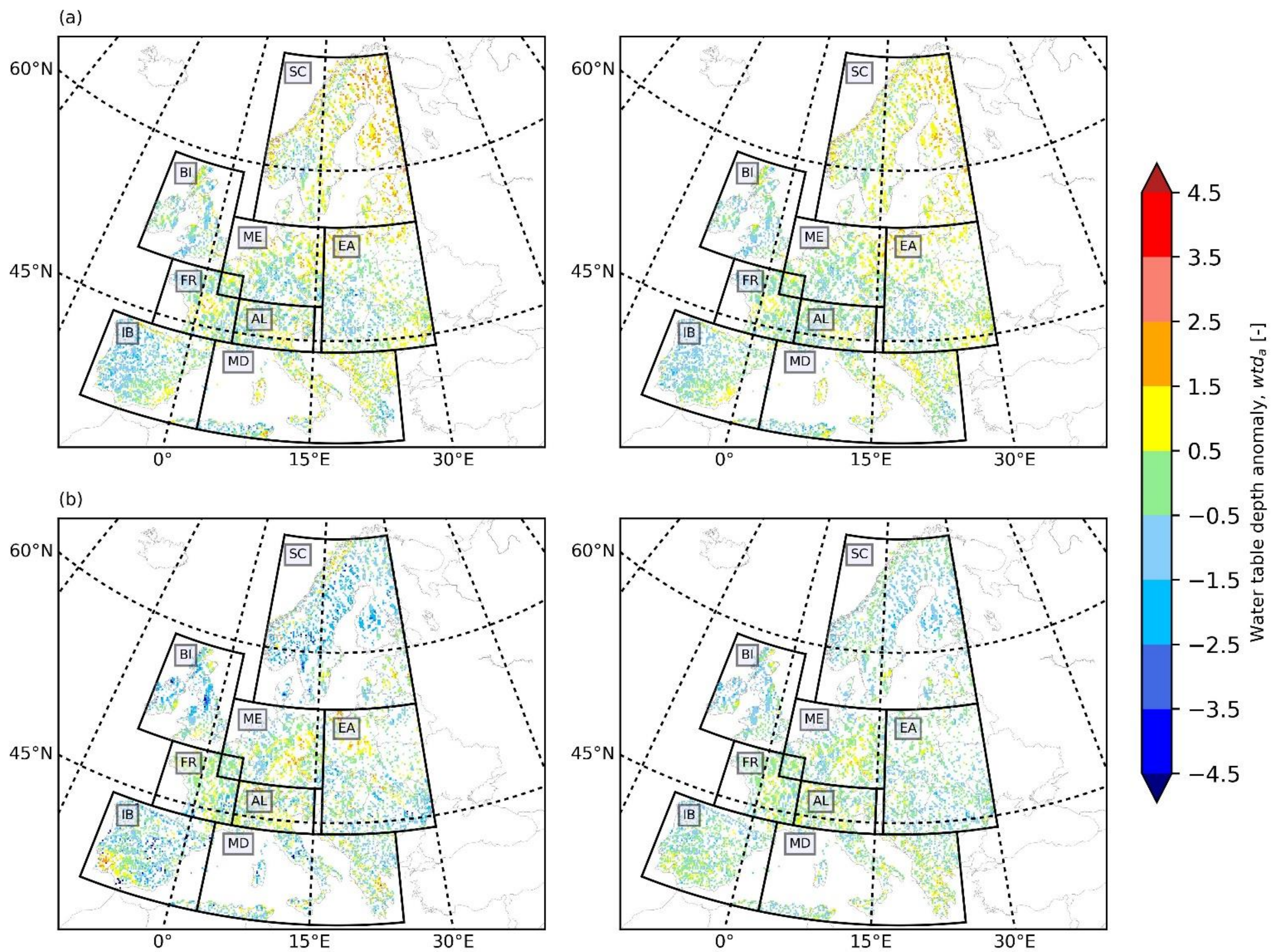

Figure B1: European wtd_a maps for (a) July 2003; (b) July 2015, derived from the TSMP-G2A data set (left) and results from LSTM networks (right).

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.
Hydrology and Earth System Sciences
Open Access
Discussions

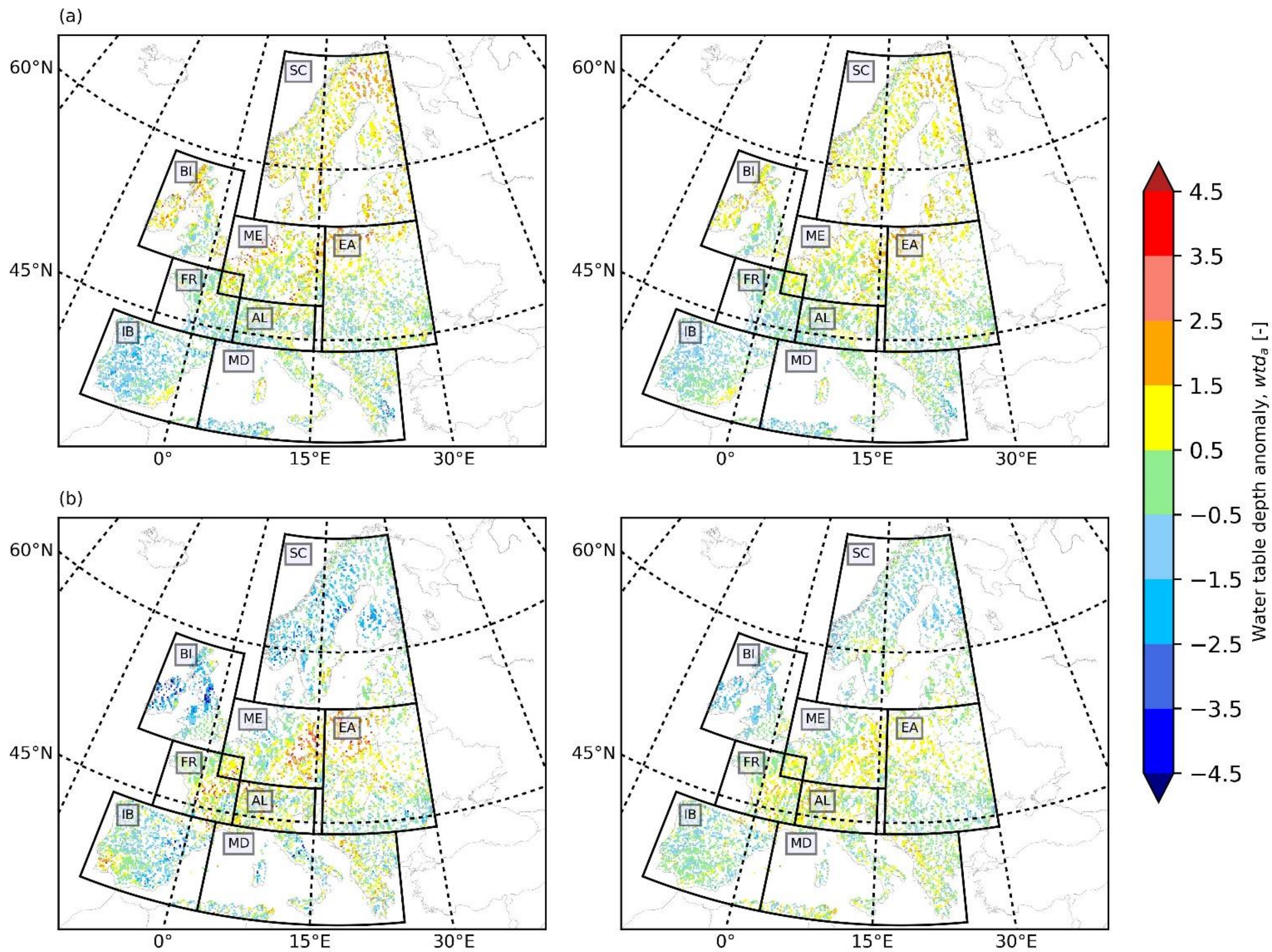

Figure B2: European wtd_a maps for (a) December 2003; (b) December 2015, derived from the TSMP-G2A data set (left) and results from LSTM networks (right).

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

Table C1: Pixel characteristics in the XWT analysis (Pixels 4-6).

	Performance combination	Region	Yearly averaged water table depth, *wtd* [m]	Yearly averaged evapotranspiration, *ET* [mm]	Yearly averaged soil moisture, θ [m^3m^{-3}]	Yearly average snow water equivalent, S_w [mm]
Pixel 4	C1	FR	1.01	418.39	0.29	0.0
Pixel 5	C2	IB	6.15	153.92	0.16	0.0
Pixel 6	C3	ME	22.59	213.96	0.19	11.0

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.

	Training R^2 [%]	Training α [%]	Validation R^2 [%]	Validation α [%]	Test R^2 [%]	Test α [%]
Pixel 4	86.42	68.88	66.94	26.02	76.62	84.86
Pixel 5	83.74	68.76	-0.8	5.85	-742.21	4.81
Pixel 6	-8.00	43.34	20.55	31.52	-267.45	6.51

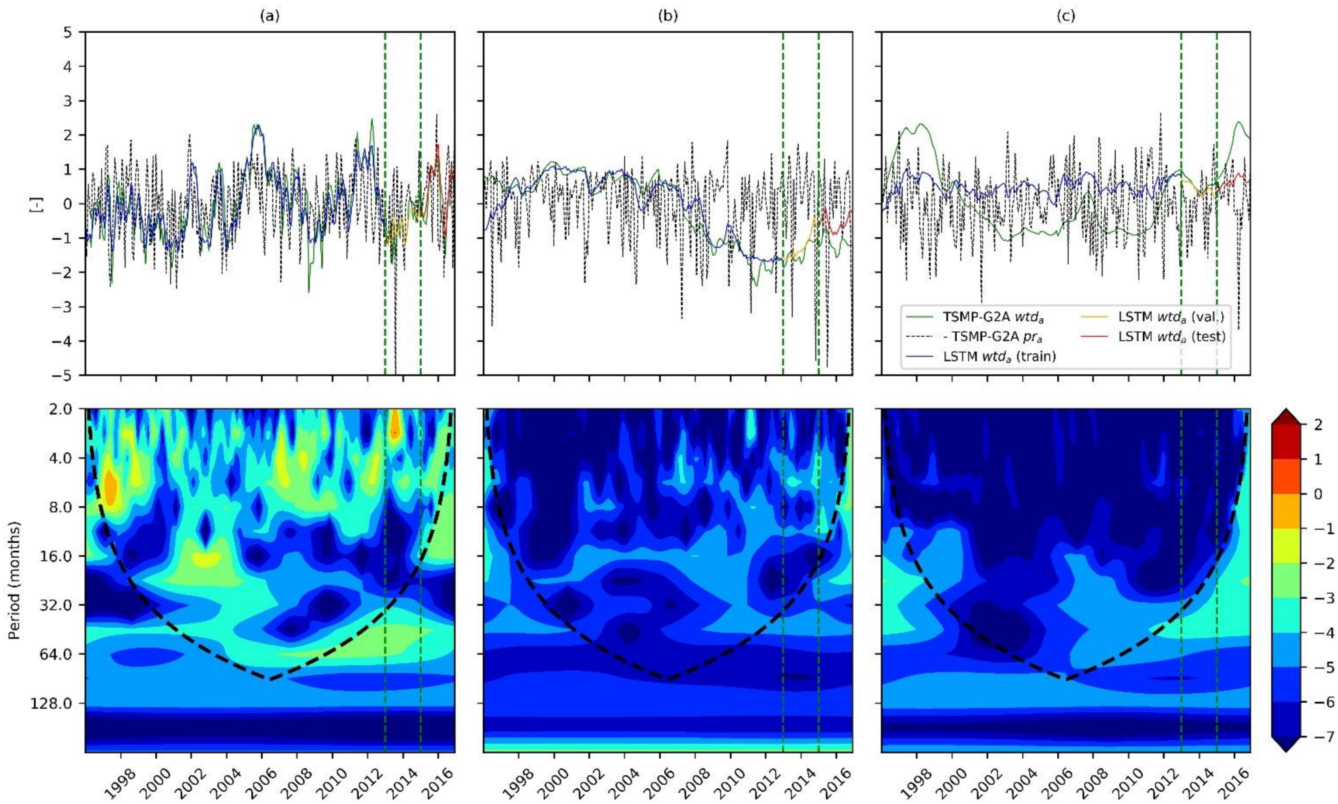

Figure C1: -TSMP-G2A *pr*$_a$, TSMP-G2A *wtd*$_a$ and LSTM *wtd*$_a$ time series (top) as well as cross-wavelet spectra for TSMP-G2A *pr*$_a$ and *wtd*$_a$ series (bottom) at (a) Pixel 4; (b) Pixel 5; (c) Pixel 6. The lines have the same definitions as Fig. 9.

Author contributions. YM and SK conceived and designed the experiments. YM conducted all the experiments and analyzed the results with feedback from CM, BB and SK. YM prepared the manuscript with contributions from all co-authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was supported by the European Commission HORIZON 2020 Program ERA-PLANET/GEOEssential project [grant number: 689443]. The authors gratefully acknowledge the computing time granted through JARA-HPC and the VSR commission on the supercomputer JUWELS at the Research Center Jülich.

References

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.

Adamowski, J. F.: River flow forecasting using wavelet and cross-wavelet transform models, Hydrol. Process., 22(25), 4877–4891, doi:10.1002/hyp.7107, 2008.

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.

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.

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.

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.

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.

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.

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

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.

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.

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.

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.: Bagging and Other Ensemble Methods, in Deep learning, MIT Press, Cambridge, Massachusetts, 250–251, 2017.

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.

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.

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

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, n.d.

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

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.

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.

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.

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.

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.

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.

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.

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.

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.

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.

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.: Generalization, Overfitting, and Underfitting, in Introduction to machine learning with Python: A GUIDE FOR DATA SCIENTISTS, O'Reilly Media, Inc., Sebastopol., 30, 2017.

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.

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.

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.

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.

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.

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.
Hydrology and Earth System Sciences
Discussions

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.

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, 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.

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.

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.

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.

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.

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

https://doi.org/10.5194/hess-2020-382
Preprint. Discussion started: 24 August 2020
© Author(s) 2020. CC BY 4.0 License.

Hydrology and Earth System Sciences Discussions

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.

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.