Journal of Hydrology: Regional Studies 47 (2023) 101416

Contents lists available at ScienceDirect

Journal of Hydrology: Regional Studies

journal homepage: www.elsevier.com/locate/ejrh

Modeling hydraulic heads with impulse response functions in different environmental settings of the Baltic countries

Marta Jemeljanova [a,b,*], Raoul A. Collenteur [c], Alexander Kmoch [b], Jānis Bikše [a], Konrāds Popovs [a], Andis Kalvāns [a]

[a] Faculty of Geography and Earth Sciences, University of Latvia, 1 Jelgavas Street, Riga, LV-1004, Latvia
[b] Department of Geography, Institute of Ecology and Earth Sciences, University of Tartu, Vanemuise 46, Tartu, 51003, Estonia
[c] Department Water Resources and Drinking Water (W+T), Eawag, Überlandstr. 133, Dübendorf, CH-8600, Switzerland

ARTICLE INFO

Dataset link: http://dx.doi.org/10.5281/zenodo.7890699

Keywords:
Time series models
IRF
Hydraulic head
SHAP values
Environmental variables

ABSTRACT

Study region: The Baltic countries (Estonia, Latvia, Lithuania), North Eastern Europe.
Study focus: Time series models are a convenient tool for modeling hydraulic head time series due to their parsimonious parameterization and quick calculation time. However, it is unclear how the model structure and the outcome relate to the respective site characteristics. Time series modeling with impulse response functions was used to model the head time series. The correlations between the model performance and the environmental settings and between the model parameters and the environmental settings were analyzed using Spearman rank correlation and Random Forest regression analyses.
New hydrological insights: From the 460 analyzed head time series, 145 were modeled with satisfactory goodness-of-fit in the calibration period, using only precipitation, potential evaporation, and temperature as model inputs. This number decreased to 68 time series with a satisfactory fit in the validation period. Including additional drivers in wells where a substantial influence is characteristic could improve modeling results. The model results suggest that snow processes are affecting head dynamics in many wells in the Baltic countries. Most correlations between the model performance and the environmental setting were observed with the geological and climatic setting of the site. Models performed best in monitoring wells with shallow groundwater and lower precipitation seasonality.

1. Introduction

Groundwater is an essential part of the hydrological cycle providing a steady water supply to ecosystems (Kløve et al., 2011) and for human consumption (Kitterød et al., 2022). Replenishment (recharge) and depletion (discharge, abstraction) of the groundwater resources have a strong seasonal character driven by changing weather patterns. As a result, strong seasonal and intra-seasonal fluctuations of groundwater hydraulic heads (hereafter 'heads') are observed worldwide (Jasechko et al., 2014). Head dynamics are affected by various environmental factors, including climate, land use and cover, vegetation, geological setting, and artificial abstraction. Many environmental processes are controlled by local fluctuations of heads, which are as diverse as greenhouse

* Corresponding author at: Faculty of Geography and Earth Sciences, University of Latvia, 1 Jelgavas Street, Riga, LV-1004, Latvia.
E-mail address: marta.jemeljanova@ut.ee (M. Jemeljanova).

https://doi.org/10.1016/j.ejrh.2023.101416
Received 9 December 2022; Received in revised form 8 May 2023; Accepted 9 May 2023
Available online 22 May 2023

gas emissions from soils (Evans et al., 2021; Mander et al., 2022), leaching and attenuation of fertilizers and pesticides from farmlands (Baran et al., 2021; Marsala et al., 2020) or river base flow (Hendriks et al., 2014). To gain valuable insights into the aforementioned processes, a robust and simple method to model heads would be beneficial.

The models used to simulate heads vary in their level of complexity, both in spatial and temporal discretization, and the level of represented physical processes. Relatively simple one-dimensional lumped parameter models use time series of input drivers and calibration targets without considering the physical characteristics of the studied location (Devia et al., 2015). Since only a point in space is modeled (e.g., head fluctuations in monitoring well in the study of Mackay et al., 2014), spatial conclusions drawn from individual instances of such models may not be valid. Although this is a substantial drawback of the modeling approach as compared to distributed process-based models, lumped parameter models require a lower amount of input data, a shorter calibration time, and a lower effort for model development, while generally good model fits are obtained (Bakker and Schaars, 2019).

Time series models using impulse response functions (IRF), as proposed by von Asmuth et al. (2002), are lumped parameter models that translate precipitation and other drivers into the output, the head. It has been shown that the use of such models is a simple and effective method for predicting head fluctuations (e.g., Marchant and Bloomfield, 2018). The impulse response function characterizes head response to an impulse of a driver, such as a precipitation event (von Asmuth et al., 2001). The function can take various shapes and thus is capable of describing a broad range of possible relationships between the impulse (precipitation or another driver) and the response (the head) (Bakker et al., 2008, 2007; von Asmuth et al., 2002). Being a lumped parameter model, a model instance is created for each head time series individually (Collenteur et al., 2019).

Time series models with impulse response functions have been applied to a variety of problems, ranging from gap filling of time series (Peterson and Western, 2018), hindcasting (Babre et al., 2022), to recharge estimation (Collenteur et al., 2021; Hocking and Kelly, 2016). Another common application is to quantify the effects of different drivers on the heads (Brakenhoff et al., 2022; Shapoori et al., 2015b; van Dijk et al., 2020). These models can be calibrated to head time series with irregular measurements (Bierkens et al., 1999), as is the case for most historical groundwater level observations (Retike et al., 2022).

Careful consideration should be given to selecting an appropriate model structure. A good resulting fit of time series models indicates that the head fluctuations can be explained well by the explanatory series. A low fit, on the other hand, can be caused by missing or non-representative input drivers, or model structural errors, among other reasons (Zaadnoordijk et al., 2019). A linear precipitation excess model calculates recharge as the difference between precipitation and evaporation. A nonlinear recharge model, on the other hand, also considers moisture retention in the soil and the limitation of actual evaporation by soil-moisture availability. This can improve the simulation of the heads (Peterson and Western, 2014; Collenteur et al., 2021). Recently, Collenteur (2022) extended a nonlinear recharge model with a degree-day snow model (e.g., Kavetski and Kuczera, 2007) to improve the model fit at sites where snowfall occurs regularly.

The relationships between the site characteristics (e.g., depth to water table) and the model fit or model parameters, respectively, are currently not well understood. IRF-based models have mainly been used as a tool in studies that characterize links between the environment and groundwater variability (e.g., Hocking and Kelly, 2016; Lu et al., 2021; van Dijk et al., 2020) or that describe similarities between some environmental factors and the groundwater (e.g., Manzione et al., 2017; Long and Mahler, 2013). The links between model performance and the study site characteristics have not been extensively described beyond the influence of individual factors (e.g., the thickness of unsaturated zone, Zaadnoordijk et al., 2019), and aquifer properties (Obergfell et al., 2013; Shapoori et al., 2015a). To the best knowledge of the authors, correlations between the IRF parameters and the environmental variables were only systematically explored in a study of Switzerland with a limited number of wells (Collenteur, 2022). The study showed that such relations may exist, but also concluded that additional research with more data is required.

The aim of this study is to systematically analyze and better understand the relationships between (1) the environmental variables and the model fit, and (2) the environmental variables and the model parameter values. To achieve these aims, the following steps were taken. In the first step head time series from the Baltic countries were modeled using four different model structures with increasing complexity. In the second step, the differences in goodness-of-fit between the model structures were compared using the Nash Sutcliffe efficiency (NSE). In the third step, the correlation between the model fit (NSE) and the environmental variables was explored with the Spearman rank correlation analysis. In the fourth step, Random Forest (RF) regression models were built with environmental variables as predictors and NSE values as the prediction target, to gain additional insights into the variables that may explain where a model structure works well and where not. To explore the correlations between the environmental variables and the parameters determining the shape of IRF steps 3 to 5 were repeated using the parameters describing the IRF separately.

2. Material and methods

2.1. Study area

The Baltic states Estonia, Latvia, and Lithuania are located in the northeast of Europe and span 175 117 km^2 (Kriauciuniene et al., 2012). They belong to the snow climate, fully humid, with warm summers (Dfb) Köppen–Geiger climate classification type (Beck et al., 2018; Kottek et al., 2006). Locally, the climate varies between maritime and continental climates in the west–east direction with additional variations in the north–south direction and based on orography, where a lower temperature and higher precipitation are characteristic in the highlands (Bethere et al., 2017). The mean annual precipitation ranges between 600 mm and over 800 mm per year (Jaagus et al., 2010). The highest elevation point is just over 300 m a.s.l., and S-SW winds prevail (Pogumirskis et al., 2021).

2.2. Hydrogeological setting

The study area is located in the Baltic Artesian Basin, and the three Baltic countries cover most of its terrestrial part. The Baltic Artesian Basin is a multi-layered sedimentary basin with vast amounts of groundwater (Kitterød et al., 2022). It consists of many aquifers forming three principal zones according to water exchange intensity: stagnant, passive, and active water exchange zones, the latter of which is predominantly used for water supply (Lukševičs et al., 2012). Aquifers typically consist of weakly cemented terrigenous and carbonate rocks, and evaporites are less common. The principal zones are separated from each other by regional aquitards which are formed mostly of marls and clays (Kitterød et al., 2022).

The active water exchange zone is formed by Cambrian-Vendian (Northern Estonia), Silurian-Ordovician (central and Western Estonia), Upper-Middle Devonian (Southern Estonia, Latvia, Lithuania), and Cenozoic-Mesozoic (Lithuania) aquifer systems (Kitterød et al., 2022). Quaternary deposits (glacial, glaciolacustrine, glaciofluvial, and marine sediments) are very heterogeneous and form the upper part of the active water exchange zone over bedrock aquifers with a thickness of a few meters in the northwestern part of the territory and the lowlands, up to 200 meters in the uplands and more than 300 meters in the buried valleys (Popovs et al., 2015, 2022). Although less so than the deeper bedrock aquifers, Quaternary groundwater is also used for centralized and decentralized water supply and is a major source of water for groundwater-dependent ecosystems (Kalvāns et al., 2021; Terasmaa et al., 2020).

2.3. Data collection and preparation

2.3.1. Hydraulic head data

The head data were obtained from various organizations responsible for maintaining head monitoring networks in the Baltic states: the Republic of Estonia Environment Agency, the Latvian Environment, Geology, and Meteorology Center, and the Lithuanian Geological Survey. The data set consists of 1671 time series of heads and the metadata measured in each of them. The metadata includes information on the respective station (if applicable), coordinates, the elevation of the monitoring well, the respective aquifer, well screen, and depth to groundwater information. Varying start and end dates and data gaps with different lengths are present in the data. The monitoring frequency of heads varied from a few measurements per year to twice a day measurements. The automatic data logging (two measurements per day) started after 2004 in Lithuania and after 2009 in Estonia and Latvia.

The raw head measurements were pre-processed using the four-eye principle and expert judgment following the methodology outlined in Retike et al. (2022). We deleted visible outliers and corrected shifts and systematic errors, i.e. errors caused by automatic logger problems, short-term abstraction impacts, human-introduced, and other miscellaneous errors. The time series used for the analysis were selected based on two criteria. First, the required length of the time series was set to 15 years, with no more than 12 data gaps, where a data gap corresponded to a month where there were no observations. If the length of the time series was longer, the most recent 15 years of data were selected. It ensures that the time series contain more observations since generally, the measuring frequency was higher in the most recent years. The data for periods where two measurements per day were taken was averaged to daily data as part of the pre-processing step.

The second criterion was that the time series do not have a substantial trend during the calibration period, chosen as the first ten years of the 15-year time series. The trend analysis was performed by calculating Sen's slopes (Sen, 1968). Before the calculation, time series were aggregated by the mean over 3-month time steps. Such a value was chosen because a regular time step between measurements was necessary to perform the trend analysis. Additionally, this aggregation step was chosen to balance between decreasing the noise that might impact the calculated trend and maintaining enough entries in the time series for the test, given the varying frequencies in the head data. The head values were re-scaled in the range [0; 1] by subtracting the mean and dividing over the range of the values.

We argue that small trends in the head data may be explained by trends in the explanatory time series. Selecting only time series with a slope of zero would exclude head time series that may still be explained well with the input data used in this study. Therefore, the choice for the threshold-value was made semi-quantitatively. Its value was determined as the elbow point in the histogram of all slope values (bins=20). The elbow point is the inflection point where the curvature changes, and it was determined visually. An additional visual evaluation was subsequently performed on a subset of the series below the determined threshold to ensure that it corresponds to series that are approximately stationary. Time series with a median slope of less than or equal to 0.07 (in absolute values) per 3 months were selected for further modeling.

The resulting data set after the different pre-processing steps is as follows. Of the original data set with 1671 time series, 735 time series cover a minimum 15-year period. Of these 735 time series, 460 had a slope smaller or equivalent to the chosen slope threshold (0.07 over 3 months for the re-scaled head series) in the calibration period. These 460 time series are used for further analysis in the remainder of this paper. The locations of the selected monitoring wells are shown in Fig. 1.

2.3.2. Meteorological data

The daily precipitation sum and the mean, maximum, and minimum temperature data were obtained from the European Climate Assessment & Dataset E-OBS gridded data ensemble (v.25.0e, resolution 0.1 °) (Cornes et al., 2018). The potential evapotranspiration (PET) was calculated using the Hargreaves–Samani equation (Hargreaves and Samani, 1985) from the temperature data. The time series of all of these drivers are available from the year 1953 to 2021.

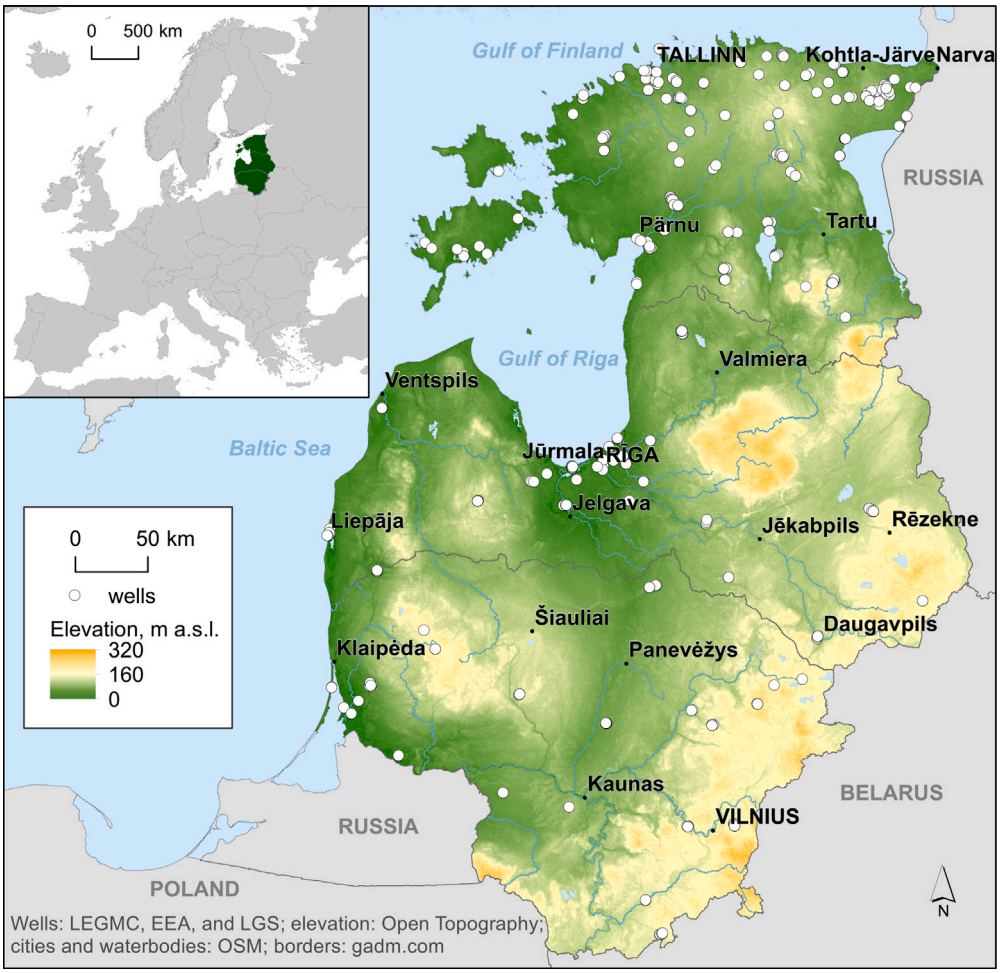

Fig. 1. The placement of the monitoring wells selected for modeling across the Baltic countries. The inset map shows the location of the three countries within Europe.

2.3.3. Environmental variables

To determine the correlations between the environmental setting of the monitoring well, the model fit, and the parameters of the impulse response functions, a set of environmental variables were derived from climatic, topographic, geologic, and boundary data following Haaf et al. (2020). This data set was further extended with land cover variables. In total, 11 climatic, 15 geological, 7 boundary, 5 land cover, and 8 topographic variables were computed, summarized in Table 1. The histograms of the environmental variables can be seen in Fig. S2 of the Supplement.

The geological variables were prepared using the lithology and geological cross-section data from Popovs et al. (2015) and Virbulis et al. (2013). The lithology data contained 12 unique classes. The dataset contained information on the geological indices and the depth of both the top and the bottom of the geological layers present in each of the cross-sections.

The boundary variables were calculated using a digital elevation model (DEM) with a 10 m resolution compiled from respective 10 m DEMs from the Republic of Estonia Land Board (Republic of Estonia Land Board, 2022), the Latvian Geospatial information agency (Latvian Geospatial information agency, 2022), the Lithuanian Geological Survey (provided upon request) and water body data from *Open Street Map* (Geofabrik GmbH and Contributors of Open Street Map, 2018). The topography parameters were calculated for 500 m buffers around the monitoring wells using the compiled DEM.

The climatic variables were calculated using daily mean temperature and daily precipitation from the E-OBS data for the respective 10-year period of calibration using the closest cell center values for each monitoring well.

The land cover variables were extracted over a 500 m buffer around the monitoring well from the Copernicus CORINE Land Cover (CLC) data (European Environment Agency, 2018), version 2018, to the first level (e.g., group 1. Artificial Surfaces) and the area within the buffer was summarized.

Table 1
Environmental variables used in the study, adapted from Haaf et al. (2020). Variable groups: geological (G), topographic (T), boundary (B), climatic (C), and land cover (LC). Variables marked with * were removed due to multicollinearity (See Fig. S1 in Supplement).

Group	Variable	Description	Value range	Units
G	A_thickness	Aquifer thickness	0.5; 197.4	m
	Depth_to_GW	Depth to groundwater head	−0.55; 62.5	m
	Screen_Upp	Upper level of well screen	0.0; 248.0	m
	Screen_Low*	Lower level of well screen	1.0; 267.4	m
	Aquifer_Depth*	Thickness of unsaturated zone	1.0; 267.4	m
	GW_level_mean*	Elevation of mean groundwater level	−28.6; 163.0	m a.s.l.
	Silt_loam_m	Cumulative thickness of silt and loam within profile	0.0; 60.0	m
	Peat_m	Cumulative thickness of peat within profile	0.0; 11.0	m
	Clay_Silt_m	Cumulative thickness of sand within profile	0.0; 99.4	m
	Marl_m	Cumulative thickness of marl within profile	0.0; 127.2	m
	Dolomite_m	Cumulative thickness of dolomite within profile	0.0; 178.4	m
	Limestone_m	Cumulative thickness of limestone within profile	0.0; 127.2	m
	Sandstone_siltstone_m	Cumulative thickness of sandstone and siltstone within profile	0.0; 170.0	m
	Gravel_m	Cumulative thickness of gravel within profile	0.0; 16.7	m
	Sand_m	Cumulative thickness of sand within profile	0.0; 69.0	m
T	arbel	Area below monitoring well (OW): percentage of area below OW height	0.0; 99.4	%
	hgthps	Position of OW relative to the height range	12.8; 100.0	%
	cncv	Concavity: percentage of area with concave structure	0.0; 56.5	%
	cnvx	Convexity: percentage of area with convex structure	0.0; 53.0%	
	cnvr	Mean convergence index	−7.6; 1.4	–
	slp*	Mean slope	0.0; 7.7	–
	twi	Mean value of topographic wetness index	7.1; 10.8	–
	ctchr	Catchment area/flow accumulation: maximum value of catchment area/flow accumulation	0.04; 597.5	m2/max number of grid cells contributing to each grid cell within radius
B	dist_boundary	Estimated distance from the GW OW to the closest segment of the outer aquifer boundary	6.3; 3120.0	m
	dist_stream	Estimated distance from the GW OW to the nearest stream (first- and second-order rivers)	4.0; 9567.5	m
	height_ow	Estimated elevation of GW OW	0.2; 206.6	m a.s.l.
	height_stream*	Estimated elevation of closest stream segment	0.0; 186.4	m a.s.l.
	height_to_stream	Meters that the stream is below/above the GW OW (gradient to stream)	−7.08; 41.5	m
	height_boundary*	Estimated elevation of closest segment of outer aquifer boundary	1.1; 210.0	m a.s.l.
	height_to_boundary	Meters that the aquifer boundary is below/above the OW (gradient to upper boundary)	−72.8; 81.9	m
C	P_avg	Mean annual precipitation	507.2; 902.2	mm
	T_avg	Mean annual temperature	4.1; 8.1	°C
	PE_avg	Mean annual potential evaporation transpiration, calculated with the Hargreaves–Samani equation	569.6; 738.8	mm
	AI	Aridity Index	0.8; 1.4	–
	meanPcoldest*	Mean precipitation in the coldest quarter of the year	22.4; 47.2	mm
	meanPwarmest*	Mean precipitation in the warmest quarter of the year	59.3; 108.1	mm
	meanPdriest	Mean precipitation in the driest quarter of the year	22.4; 47.2	mm
	meanPwettest	Mean precipitation in the wettest quarter of the year	59.3; 108.1	mm
	ratioPcoldwarm*	meanPcoldest divided by meanPwarmest	0.3; 0.7	–
	ratioPdrywett*	meanPdriest divided by meanPwettest	0.3; 0.7	–
	SI	Seasonality index of precipitation	0.1; 0.4	–
LC	Artificial	Fraction of area of the CLC class Artificial surfaces	0.0; 1.0	–
	Agricultural	Fraction of area of the CLC class Agricultural areas	0.0; 1.0	–
	Forest	Fraction of area of the CLC class Forest and seminatural areas	0.0; 1.0	–
	Wetlands	Fraction of area of the CLC class Wetlands	0.0; 1.0	–
	Water_bodies	Fraction of area of the CLC class Waterbodies	0.0; 1.0	–

2.4. Methods

2.4.1. Time series modeling using impulse response functions

Time series models using impulse response functions (IRF) are used to model the head time series (von Asmuth et al., 2002). Predefined response functions are used to characterize the head response to a unit pulse of a driver (e.g., precipitation). In this study, daily precipitation, potential evaporation, and mean temperature data are used as drivers. The head at each time step is modeled as a superposition of the lagged driver impacts: the value of the driver of each lagged time step is multiplied by the impulse response

Table 2
Model structures employed in the study. The PE-model refers to the model used for precipitation and evaporation.

Abbreviation	PE-Model	Response function	Number of parameters in the IRF	Accounts for snow processes
LG	Linear	Gamma	3	No
L4	Linear	Four parameter	4	No
NLG	Nonlinear	Gamma	3	No
NLS	Nonlinear	Gamma	3	Yes

value of the respective time step and subsequently summed. The shape and size of the response function are described by only a few parameters, which are estimated by fitting the simulated head time series to the observed.

The choice for a certain IRF type depends on its capability to characterize the head response accurately. A scaled Gamma distribution function is often used to simulate the head response to different drivers (e.g., von Asmuth et al., 2008). The impulse response function is written as:

$$\theta_p(t) = A\frac{t^{n-1}}{a^n \Gamma(n)} e^{-t/a} \quad \text{for } t \geq 0 \tag{1}$$

where A, a, and n are parameters that describe the shape of the response function. Alternatively, a four-parameter function may be used (Bakker et al., 2008), for which the impulse response function is approximated as follows:

$$\theta_p(t) = At^{n-1} e^{-t/a-ab/t} \quad \text{for } t \geq 0 \tag{2}$$

where A, n, a, b are the shape parameters of the response function. For both response functions, the parameter A corresponds to the gain of the system (the magnitude of the impact), and the parameters n and a determine the shape, i.e., larger values of n and a simulate a more lagged response to an impulse of a driver. The parameter b of the four-parameter function (Eq. (2)) allows for a larger delay in the response, by shifting the response in time. The parameter values of the different impulse response functions are unknown and need to be inferred from the data.

Two approaches are applied to account for precipitation and potential evaporation in the model – a linear precipitation excess model and a nonlinear root zone model. These models are only briefly introduced here and we refer to Collenteur et al. (2021) and references therein for a more detailed explanation. The linear model assumes a linear relationship between precipitation and evaporation. The nonlinear model, on the other hand, accounts for the nonlinear response of the head to precipitation and evaporation. This nonlinear model assumes that the system can be modeled as two connecting reservoirs — an interception and root zone reservoirs, thus including short-term water retention in the soil in the calculations. Under certain conditions, a linear model combined with a more complex impulse response function can produce results similar to those of the root zone model. However, as also stated by Peterson and Western (2018), the missing soil moisture representation may introduce possible errors. If snow processes (e.g., snowfall and snowmelt) are of importance, a snow reservoir may be added on top of the nonlinear model (Collenteur, 2022).

In general it can be assumed that the model that is most appropriate for each individual monitoring well is unknown. A multi-model approach is therefore applied here, assessing the performance of four different model structures with increasing complexity to simulate the heads. These include linear and nonlinear approaches to take precipitation and evaporation into account, combined with a Gamma or four parameter function to describe the shape of the response. The models used in the study were (from the simplest to the most complex): a linear model with Gamma function (LG), a linear model with a four parameter function (L4), a nonlinear model with a Gamma function (NLG), and a nonlinear snow model with a Gamma function (NLS) (see Table 2). Different hydrological processes are represented in these models and the models have an increasing amount of parameters that need to be estimated.

2.4.2. Model calibration and evaluation

The models were calibrated by fitting the simulated heads to the observed ones. The residuals of models simulating head time series often show strong autocorrelation and are therefore first modeled with an auto-regressive noise model of order one (AR1). The resulting noise time series, which should have reduced autocorrelation compared to the raw model residuals, is then minimized using a nonlinear least-squares approach. The parameters were estimated first without the noise model and subsequently with a noise model to ensure better parameter calibration. For a thorough explanation of this calibration process, readers are referred to von Asmuth and Bierkens (2005) and Collenteur et al. (2021). The models were calibrated over a 10 years period. It should be noted that the start and end dates of calibration for each monitoring well change depending on the available head data for that well. A warmup period of 10 years was applied, and the modeling time step was 1 day. The performance of the AR1 noise model was evaluated with the non-parametric Stoffer–Toloi test to determine whether the autocorrelation is significantly different from zero on non-equidistant time series. A significance level of 0.05 was used.

The model fit was determined by a combination of an evaluation metric and visual evaluation, as suggested by Bennett et al. (2013), Legates and McCabe (1999), Moriasi et al. (2007). The Nash Sutcliffe efficiency (NSE) and the root mean squared error (RMSE) were calculated. NSE is the ratio between the variance of the residuals and the variance of the observations with resulting values in the range [-∞; 1]. A value above zero indicates that the simulation gives a better result than using the mean of the values (Gupta and Kling, 2011; Knoben et al., 2019). The RMSE values span [0; ∞] (Moriasi et al., 2015) and are in the units of measure (Legates and McCabe, 1999). Both metrics were calculated for the calibration and validation periods separately. Models

reaching the threshold of 0.65 for NSE and passing visual evaluation (does the simulation follow the trend of the observations in the calibration period) were determined as having a good fit. The RMSE values were used to additionally assess the model fit, but not used as a threshold because the amplitudes of the head series were different. Mann–Whitney U test (e.g., Nachar, 2008) was performed on the obtained NSE values over the calibration period of the different model structures to determine if the resulting difference is statistically significant.

2.4.3. Assessing correlations between model performance, model parameterization, and the environmental variables

To investigate if changes in the model performance and calibrated parameter values correspond to those of the independent environmental variables, correlation and regression analyses were performed. The aim of these analyses is twofold: (1) to determine which model structures work best in which environmental settings as quantified by the environmental variables, and (2) to investigate how the parameters of the response function are related to the environmental settings. For the first purpose, the NSE value of a model during the calibration period is used here to quantify the model performance.

For the analysis between the model performance and the environmental variables, the NSE values during the calibration period of all models irrespective of the model fit were used. For the analysis concerning the parameters A, n, and a of the IRF, however, the analysis was performed on the parameter values of the models that reached a good fit for all model structures. Using only the time series where all model structures passed the threshold ensures that the models could reasonably quantify the head fluctuations with the calibrated parameters. For both analyses, an equal subset of the time series was used for all four model structures, thus the observed strength of relations could be compared between model structures.

To determine if the NSE or parameter values are related to the value of environmental variables, a non-parametric Spearman rank correlation analysis was performed. This was done for each environmental variable and the NSE or IRF parameter values for each model structure separately. Both the correlation coefficients and their statistical significance were calculated. The significance level used to determine if the correlations are significant was set to 0.05.

Subsequently, a Random Forest regression was applied together with a SHapely Additive exPlanations (SHAP) values analysis and plotting partial dependence. Random Forest is a widely used non-parametric machine learning approach for regression problems. The prediction is obtained by creating decision trees from randomly selected inputs and averaging their outcome. The advantages of this regression approach include robustness to overfitting (Breiman, 2001) and its explainability. The concept of SHAP values has recently been demonstrated to be useful to quantify the impact of each explanatory variable on the prediction in machine learning models, e.g., stream nutrients in Virro et al. (2022) or groundwater levels in Wunsch et al. (2022). The SHAP values can be calculated both globally (dataset-wide) and locally (individual instances) and reveal the direction of the association (a positive or negative contribution) (Lipovetsky and Conklin, 2001). The values are in the units of the prediction target, and therefore simple to interpret (Molnar, 2022). For Random Forest models, the purpose-adjusted TreeExplainer was used to calculate the SHAP values (Lundberg et al., 2020). The SHAP value analysis is often combined with partial dependence plots. The plots visualize the contribution of an explanatory variable on the prediction target after the average impacts of other variables have been removed (the marginal effect) (Friedman and Meulman, 2003). The plots reveal if the association is linear, non-linear, or monotonic (Molnar, 2022). For further details, the readers are referred to Friedman (2001).

The NSE and the IRF parameters of each of the model structures (LG, L4, NLG, and NLS) were analyzed separately using the respective environmental variables as the independent variables and the NSE or parameter values as the prediction target. Since the RF model does not accept empty input values, the monitoring wells with at least one missing environmental variable value were eliminated before the regression analysis. The default Random Forest parameters (n_estimators = 100, max_depth = None, min_samples_split = 2) were used. Both the average and individual SHAP values were calculated. The partial dependence was plotted for 10 variables with the highest influence.

3. Results

3.1. Overall model performances

Each of the 460 time series selected for the analysis was modeled with the four different model structures. In total, 1840 models were created. Of the 1840 models, 334 models (corresponding to 145 out of 460 unique monitoring wells) had a good fit in the calibration period, i.e. they reached or exceeded the chosen metric threshold. The NLS model structure performed the best having the highest number of models with a good fit (34% of the models, respectively) (Table 3). The NLG model structure scored the lowest with 20% of models with a good fit.

Of the 334 models, 131 reached the metric threshold in the validation period (68 unique monitoring wells). The highest count of models with a good fit was again for the NLS model (31%), with the LG, L4, and NLG model structures all achieving lower numbers. A visual evaluation was performed on the hydrographs of the models that exceeded the goodness-of-fit thresholds in the validation period. No models were discarded, showing that the goodness-of-fit metric and its threshold sufficed as an evaluation in this study.

Fig. 2 shows boxplots of the goodness-of-fit metrics for all four model structures for both the calibration (upper row) and the validation period (lower row). The median NSE of the calibration period varied between 0.45 and 0.52 for the four model structures. This indicates that the models have a notable difference in modeling performance. The NLS model structure had the highest average NSE value, and the NLG model structure scored the lowest. The Mann–Whitney U test revealed that the difference in performance between both model structures is statistically significant. For any other combination of the model structures, there was no significant difference.

Table 3
Number of wells (out of 460) that reached or exceeded the NSE metric threshold of 0.65 for the respective calibration (10 years) and validation (most recent 5 years) periods. The periods of calibration and validation were different and depended on the available data.

Model structure	Calibration	Validation
LG	75	33
L4	78	27
NLG	68	30
NLS	113	41

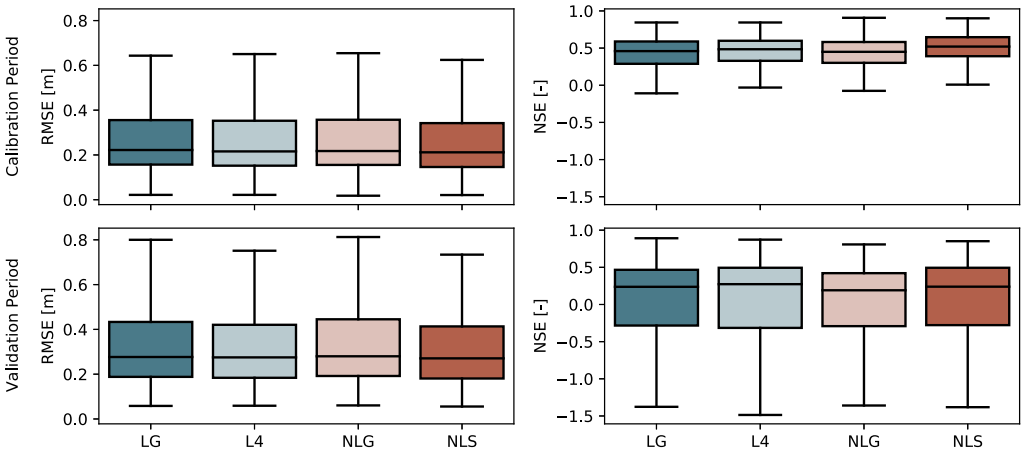

Fig. 2. Median metrics for the calibration and validation periods for each model structure (outliers removed).

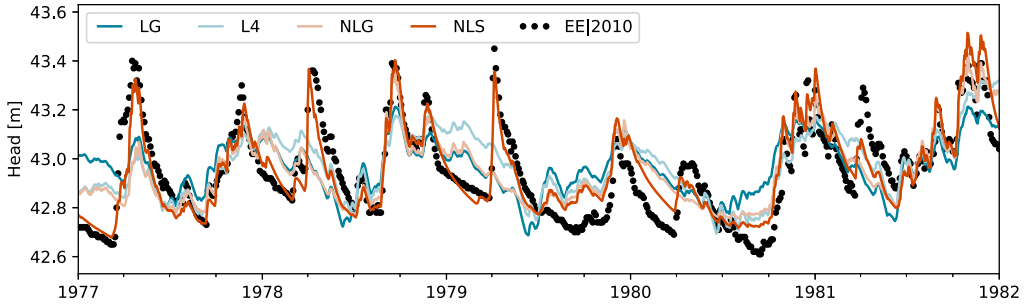

Fig. 3. Comparison of the modeling results from different model structures (monitoring well 2010, Estonia). The observed head data are plotted with black dots.

The RMSE values are in the same units as the studied variable (in this study, in meters). All of the model structures had similar RMSE values in the calibration period, with a difference of 1 cm between medians and they varied between 21 and 22 cm. The median amplitude (the maximum diversion from the mean) of the time series was 90 cm, therefore the obtained RMSE, compared against the median amplitude of the time series, indicates that the heads were generally simulated well.

In the validation period, all median metrics were worse than those in the calibration period. This indicates that less of the groundwater level variability in that period was captured compared to the calibration period. The L4 model structure had the highest NSE median value and the NLG model structure scored the lowest. However, none of the differences between model structures were statistically significant. Like in the calibration period, all model structures had very similar median RMSE values (0.27 and 0.28), and they had increased by ±6cm as compared to the calibration period. Many of the models showed particularly low performance during the validation period (i.e., NSE below zero). A visual inspection of the results showed that these low fits in the validation period were caused by (strong) trends in the head data during that period, which was not checked for the presence of trends.

We illustrate the results of model performance based on the structure with an example of well 2010 in Estonia (see Fig. 3). The NLS model structure performed best in replicating the extremes, especially in the colder months of the year. From the remaining three model structures, the NLG performed the best at replicating both peaks and lows, while both linear models performed the poorest. The obtained NSE values for model structures are similarly ordered. The NLS model structure had NSE over 0.7 for both calibration and validation periods, followed by the NLG and L4 (NSE between 0.4-0.5). The LG performed the poorest with NSE values of 0.25 and 0.3 in the calibration and validation periods, respectively.

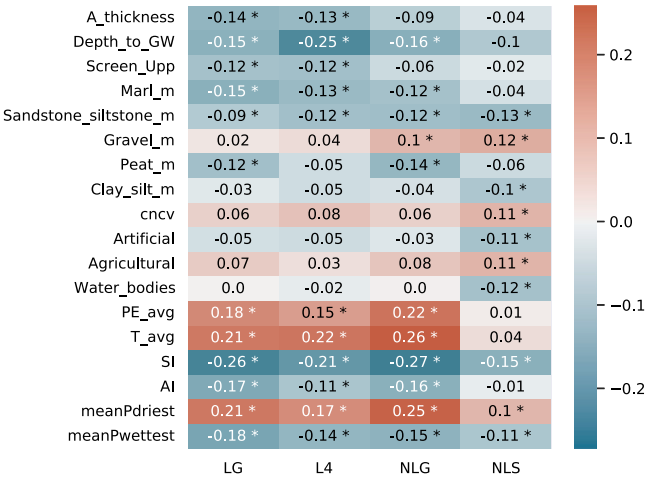

Fig. 4. Correlation coefficients between the environmental variables and the NSE values of the model structures. Coefficients with statistical significance are marked with "*". Only variables with statistically significant correlation coefficients with at least one model structure are displayed.

A Stoffer–Toloi test was performed to evaluate the autocorrelation in the noise series. Across the model structures, no significant autocorrelation was detected for approximately 33 to 42% of the models, depending on the model structure. The NLS model resulted in the largest number of models without significant autocorrelation (196 models out of 460 models), followed by the NLG (169 models). The poorest results were obtained with both linear model structures (154 models for L4 and 158 for LG).

3.2. Correlation between environmental variables and model fit

Since the same monitoring well data was used for all model types (460 unique monitoring wells), the correlation coefficients can be directly compared between the model structures. Overall, low correlation coefficients were obtained (see Fig. 4). The model fit (described with the NSE value) most strongly correlated with various climatic variables, both those that describe the average yearly meteorological conditions (*T_avg*, *PE_avg*) and the seasonality effects (*SI*, *meanPwettest*, *meanPdriest*, *AI*). Notably, all model structures' fits had negligible correlations with the average yearly precipitation (*P_avg*) (not displayed). It indicates that the changes in model fit correspond with the respective precipitation seasonality. More specifically, a higher model fit is associated with lower precipitation seasonality for all model fits (lower seasonality index, more precipitation in the driest season, and less precipitation in the wettest season).

Additionally, the model fit of NLS had negligible correlations with those climatic variables that depend on temperature (such as *T_avg*, *PE_avg*, and *AI*). The model fit of NLS correlated less with the precipitation seasonality variables. It indicates that the NLS model structure is the most robust model structure considering the seasonality setting of the monitoring wells.

The fit of both linear models was inversely correlated to the placement of the groundwater table (described with the variables *Depth_to_GW* and *Screen_Upp*) as well as the aquifer thickness (*A_thickness*), namely more shallow groundwater level and thinner aquifer in study site corresponded to a higher model fit. A similar conclusion could apply to the NLG model structure, although a significant correlation was present only with one of the variables.

Various lithology variable correlations with the model fit indicate that the model fit positively correlates with increasing thickness of highly permeable sediments (*Gravel_m*) and negatively with increasing thickness of less permeable sediments (such as *Clay_silt_m*). However, the lack of statistical significance and negligible correlation coefficients of other lithology groups indicates that within this study, there is not enough evidence to reliably claim it.

In addition to the correlation analysis, a regression analysis was performed based on the interpretable Random Forest (RF) algorithm on the data from 336 unique monitoring wells. The 336 monitoring wells out of 460 selected for groundwater modeling correspond to those that remained after eliminating monitoring wells with missing environmental variables. The obtained model fits (R^2) varied between 0.85–0.89, indicating that the NSE for all model structures could be predicted well from the selected environmental variables. These variables were not used as input in modeling the groundwater level with the time series models, therefore the relationships obtained with RF are an indirect interpretation that, nevertheless, can reveal valuable insights. The SHAP values were used to analyze which environmental factors primarily explain the variation in NSE among the different models.

The summary SHAP plot for all model structures is visualized in Fig. 5, displayed in the order of average importance descending. The most important characteristics for prediction of model fit across all model structures were the geological (*Depth_to_GW*, *Screep_Upp*) and climatic (*SI*, *meanPwettest*, and *T_avg*) setting of the monitoring wells. While the absolute SHAP values describe the general impact of the variable on the NSE values, the summary plot allows exploring the direction of the impact.

For all model structures, both larger values of *Depth_to_GW* and *Screen_Upp* contribute negatively to the model fit, indicating that a higher fit is achieved for wells where the groundwater was more shallow. Similarly, higher values of the climatic variables

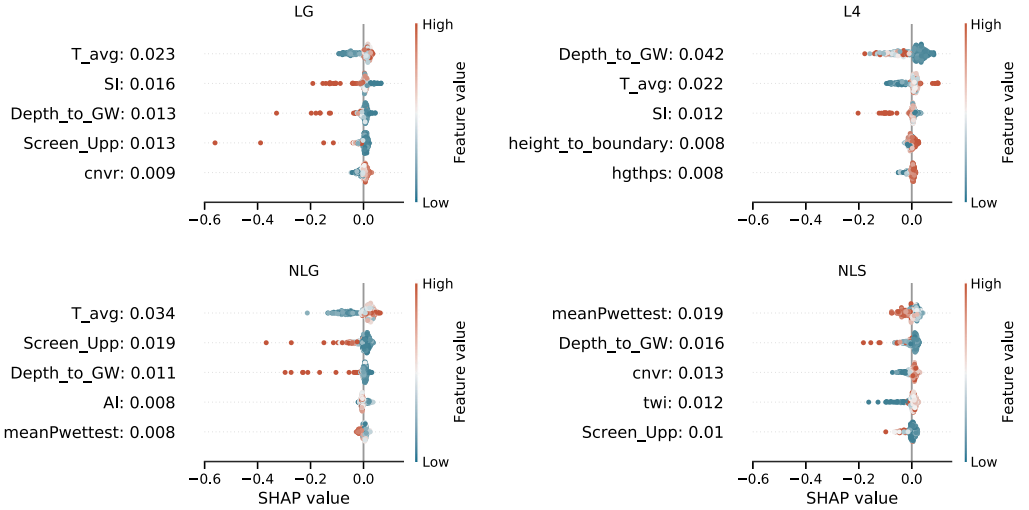

Fig. 5. SHAP summary plots for each model structure. Dots represent unique monitoring wells, and the color represents the feature value. The individual SHAP values are displayed on the *x*-axis, while the average SHAP value of each variable is displayed next to the name of the variable. A positive or negative impact on the NSE values can be seen on the *x*-axis, while the color represents the individual variable value in the range of values in the whole dataset (low or high). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

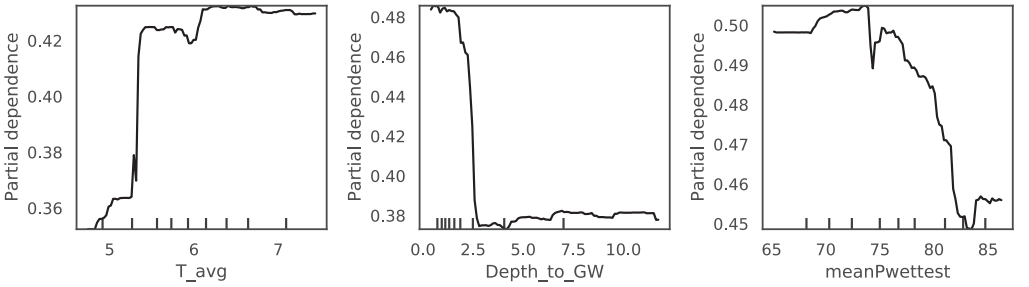

Fig. 6. Partial dependence plots for environmental variables average yearly temperature (*T_avg*, LG), depth to the groundwater table (*Depth_to_GW*, L4) and mean precipitation of the wettest quarter (*meanPwettest*, NLS). The horizontal axis shows the values of the environmental variable.

SI (Seasonality Index) for both linear models and *meanPwettest* for both nonlinear models decreased the model fit, indicating that for monitoring wells with more pronounced precipitation seasonality a lower fit of the models might be achieved. However, the climatic variable *T_avg*, present for all but NLS structure, contributed positively to the model fit.

While on average each variable had a minor influence on the NSE value (< 0.05), the influence was substantial in individual locations. For example, a deeper upper level of the screen (high values of variable *Screen_Upp*) decreased the model fit value by more than 0.5 for the LG model structure or up to 0.3 for the NLG model structure. Similarly, deeper groundwater level (higher values of the variable *Depth_to_GW*) contributed to decreasing the model fit by up to 0.2 or 0.3, depending on both the model structure and specific instance. Other variables with substantial local influence were the seasonality index for L4 and the topographic wetness index for the NLS model structure.

Across the four model structures, a larger contribution was evident towards decreasing the NSE value. The highest positive influence on the model fit, however, was by the depth of groundwater level and the mean yearly temperature for the L4 model structure. The NSE value was increased by only up to 0.1.

Selected partial dependence plots are shown in Fig. 6. Partial dependence plots provide insights into the impact of the value of the studied variable (e.g., depth to the water table) on the resulting model fit. Hence, thresholds of impact can be determined by reading the horizontal axis, where the variable values are displayed. The influence of the variable *T_avg* for the NLG model structure increases substantially from the 5.5 °C mark. The influence of the variable *Depth_to_GW* is highest until the 2 m mark, after which it drops and evens out. In the case of the variable *meanPwettest* for the NLS model structure, partial dependence increased until around the 75 mm mark and subsequently decreased. See the Supplementary material, Figs. 3 to 6 for partial dependence plots of the top 10 variables of all model structures.

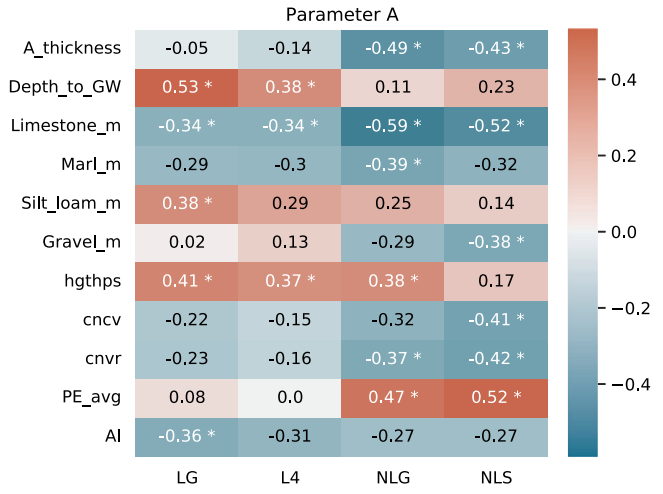

Fig. 7. Correlation coefficients between the environmental variables and the IRF parameter *A* for each model structure. Coefficients with statistical significance are marked with "*". Only variables with statistically significant correlation coefficients with at least one model structure are displayed.

3.3. Correlation between the environmental variables and the parameterization of IRF

The correlation analysis was performed for the parameters *A*, *n*, and *a* that describe the shape and size of the IRF. Weak to medium correlation coefficients were obtained. The resulting correlations for all model structures and the three parameters are shown in Figs 7, 8, and 9.

The parameter *A* that characterizes the gain of the system significantly correlates with various lithology variables (Fig. 7). With *Gravel_m*, the correlation coefficient was negative (more gravel means smaller gains), while it was positive with the variable *Silt_loam_m* which are less permeable sediments. Similarly, negative correlations were present with the variables *Limestone_m* and *Marl_m*. This indicates that the parameter *A* value is larger in locations with thicker low permeability sediments and smaller where more of less permeable sediments are present, thus, positively correlating with low permeability. Additionally, the gain parameter was larger where the groundwater table was deeper (positive correlation with *Depth_to_GW*), but the respective aquifer was thinner (negative *A_thickness*). The parameter *A* was present in monitoring wells higher in the topography, revealed by the value of *A* was larger when *hgthps* were higher and *cnvr* and *cncv* were lower. Mostly, the significant correlations were shared between the model types (e.g., *PE_avg* for both nonlinear model structures) or were individual. The reason could be that the dataset could be too small to make reliable statistical claims or there indeed are differences between the correlations based on model types. The shared correlation with *Limestone_m* indicates that different model types (linear or nonlinear) can correlate differently to some characteristics of the environmental setting.

The parameter *n* determines the time lag in the head response to the recharge or precipitation excess signals. The positive correlations with variables *Depth_to_GW* and *Screen_Upp* (Fig. 8) indicate that the value of *n* is larger in instances of a thicker unsaturated zone. Additionally, the parameter value is higher where a thicker silt layer in the profile (larger values of variable *Silt_loam_m*) is present. Similar to the parameter *A*, a higher parameter *n* value is characteristic higher in the topography (based on both *hgthps* and *arbel* variables) and in locations with more pronounced rainfall seasonality (variables *SI* and *meanPwettest*).

The parameter *a* of the IRF, similar to the parameter *n*, correlated to the depth of water table (variables *Depth_to_GW* and *Screen_Upp*). In addition, the parameter of both linear model structures (LG and L4) positively correlated with the absolute and relative elevation of the observation well (*hgthps* and *height_ow*). While within this dataset positive correlations were obtained for both nonlinear models also, there is not enough statistical evidence to claim that the correlations are significant. Similarly, a negative correlation with *Marl_m* and a positive one with *Gravel_m* could mean that the parameter *a* correlates to the permeability of sediments, however, there is not enough statistical evidence.

4. Discussion

4.1. Modeling hydraulic heads in the Baltics

Application of the time series models on the 460 head time series selected for the analysis, resulted in 145 monitoring wells where at least one model structure had a good fit with the head data. This result shows that for about a third of the monitoring wells (±30%) it is possible to simulate the heads with reasonable accuracy using relatively simple models and only a few driving forces. As such, these monitoring wells appear to observe mostly natural fluctuations with relatively little impact from anthropogenic

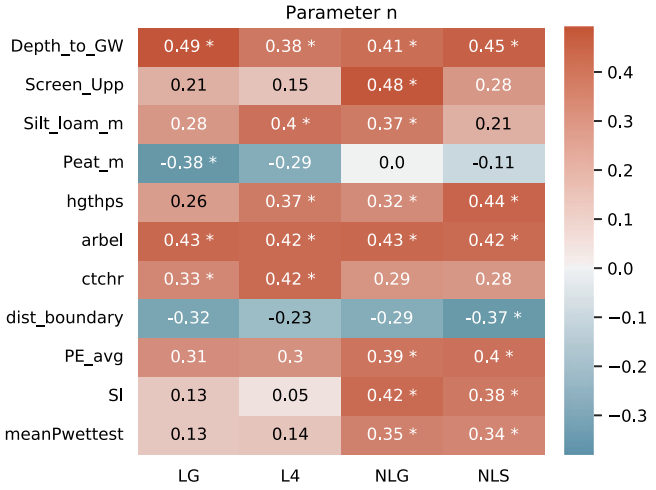

Fig. 8. Correlation coefficients between the environmental variables and the IRF parameter *n* for each model structure. Coefficients with statistical significance are marked with "*". Only variables with statistically significant correlation coefficients with at least one model structure are displayed.

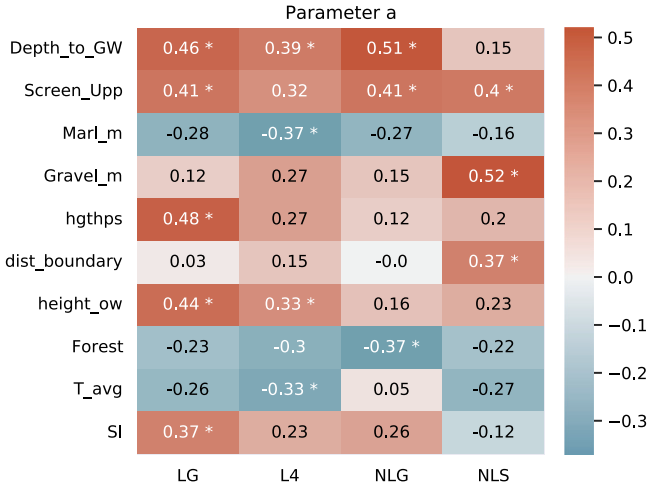

Fig. 9. Correlation coefficients between the environmental variables and the IRF parameter *a* for each model structure. Coefficients with statistical significance are marked with "*". Only variables with statistically significant correlation coefficients with at least one model structure are displayed.

influences. These results are in line with those from Zaadnoordijk et al. (2019), who found in a large-scale study in the Netherlands that less than half of the wells could be modeled satisfactorily with only precipitation and evaporation as driving forces.

For the remaining 70% of the monitoring wells in this study, the results can be interpreted as an indication that other drivers may influence the heads (e.g., pumping, surface water levels). The importance of other driving forces should not be unexpected, as monitoring networks are often established with the purpose of monitoring human influences or changes on the groundwater. As also suggested by Zaadnoordijk et al. (2019), including these additional drivers such as surface water fluctuations and pumping in the models could improve the results. Adding additional drivers to improve the modeling results is relatively easy for the type of model used in this study, as shown in von Asmuth et al. (2008) for example. Based on the results of this study, we recommend testing the use of additional input data to model the heads, such as groundwater pumping and river stage data, if such data is available.

Only 131 models reached the goodness-of-fit thresholds in the validation period. One reason for the decline in the model performance in the validation period is the presence of trends, revealed from a visual inspection, that could not be explained by the driving forces that were used. Checking for trends over both the calibration and validation period, rather than only the calibration period as was done now, could have prevented this. This would, however, lead to a higher number of wells rejected from the analysis, and not to more models with a good fit. To increase the model fit, including a linear trend in the model would

be an option to try and improve the model fit. Without information on the cause of such trends, this would, however, not lead to a better system understanding, and was therefore judged outside the scope of this study.

The most appropriate model structure to simulate the heads will depend on the environmental settings of the monitoring well. Some general patterns, however, became apparent from the modeling results. The nonlinear snow model with a Gamma response function resulted in the highest number of models with a good fit. High median metrics were also obtained for the linear model with a four parameter response function. Thus, both model structures appear to be more applicable than the other models for a large number of wells in the Baltics. The improvement associated with the NLS model structure over the other models can mostly be attributed to the inclusion of snow processes in this model, as none of the other models included this process. The linear four parameter function model can represent the delayed head response due to snow processes by shifting the response in time. Based on these results, we conclude that when modeling the heads in the Baltic countries, it is important to account for the presence of snow cover when the temperatures are below zero. This is in line with the general perception of the importance of these processes on the hydrological regime in the Baltics (e.g., Kriauciuniene et al., 2012; Jaagus et al., 2017). It is expected that these models will also perform better in other snow-dominated regions around the world.

4.2. The correlation between the environmental variables and the model fit or the model parameterization

The obtained coefficients in correlation studies were weak to moderate. However, they give an insight into similarities between the placement of the monitoring well and the modeling outcome. From the five groups of environmental variables, the strongest correlations with the model fit were found with geological and climatic variable groups. More specifically, the thickness of the unsaturated zone or the upper level of the well screen (variables *Depth_to_GW* and *Screen_Upp*, respectively) were found to be two of the most important variables in both the correlation and regression analyses, negatively impacting the model fit (e.g., lower fits for deeper water tables/ aquifers). We only speculate about possible reasons that may explain this finding here. Shallow groundwater generally responds more quickly to the precipitation signal, but as the thickness increases the head response becomes lagged. At a certain depth, it is entirely independent of the signal (Wang et al., 2022). This may decrease our ability to determine the association. Other possible reasons include inappropriate model structures, and unknown drivers such as pumping that disproportionately affect deeper aquifers (e.g., pumping for safe drinking water).

From the climatic variables, the mean precipitation of the wettest quarter (*meanPwettest*) correlated with the NSE of all model structures. It characterizes the summed precipitation of the wettest season (June, July, and August), which in this study was also the warmest season. In addition, the mean precipitation of the driest quarter (*meanPdriest*) (February, March, and April) and the seasonality index (*SI*) correlated with all model structures' NSE values and scored high in the top 10 variables in the regression analysis. The monitoring wells in the Baltics correspond to two classes of SI values: monitoring wells without precipitation seasonality (*SI* <= 0.19) and monitoring wells with a more pronounced wet season (*SI* 0.2–0.39) (Walsh and Lawler, 1981). The driest season corresponds to months with both snow cover (February, sometimes March) and spring floods (April): the months when the infiltration is highly affected by the mean temperature. The lowest correlations were present with the NSE of the NLS model structure where the mean daily temperature was used as model input. This is an additional indication that the inclusion of snow processes improves the model fit for the Baltic states. The correlations of the climatic variables indicate that the model fit is affected by the precipitation seasonality. The model fit was higher for the monitoring wells where the precipitation seasonality was less pronounced.

While similarities between model structure correlations were evident, the coefficient values varied. The NLS model structure appeared to be more robust towards the climatic impact since there were fewer statistically significant correlations with the variables of the group. There were instances where both linear models shared a significant correlation while both nonlinear models did not and vice versa. It could indicate that changing the model type to nonlinear brings more substantial changes than increasing the complexity of the IRF function. It is necessary, however, to point out that correlation coefficients without statistical significance mean that there is not enough statistical evidence to claim that the correlation is significant, but we cannot claim that there is no correlation (we cannot accept the null hypothesis, just fail to reject it) (Amrhein et al., 2019). Therefore, the results should not be taken at face value and should be evaluated in the context they represent. Further, larger-scale studies would be advisable. Thus, our conclusions are affected by our chosen dataset.

The correlation analysis performed on the models with a good fit in the validation period (not shown) revealed that the significant correlation coefficients were moderate (0.3–0.5), as opposed to weak when the analysis was performed on all models. The number of significant correlations decreased, however, most probably due to the decreased dataset. The variables correlating with model fit were from the climatic group for the LG, L4, and NLG model structures (variables *SI*, *AI*, *meanPdriest*, and *T_avg*) and topographic and boundary (*arbel* and *height_to_stream*) for the NLS model structure.

We stress that within this study, the environmental variables were not used in the IRF models as an input. The RF models were built using the independent environmental variables as the predictors and the model fit obtained from the IRF models as the prediction target. Thus, links between the environmental setting and the model fit can be explored for models that need parsimonious input. The obtained results of the SHAP value analysis reveal that depth to groundwater level, seasonality index, the upper level of the well screen, and other variables can have substantial impacts on the resulting model fit (up to 0.5 in absolute values), depending on the selected model structure and environmental settings of the monitoring well. The obtained partial dependence plots reveal that beyond the depth of 2 m, the depth of groundwater has a (close to) uniform, low impact on the modeling outcome. Similarly, the yearly average temperature has the highest impact on the modeling outcome when the temperature is higher than 5.5 degrees and the mean precipitation in the wettest quarter is lower than 73 mm. We note here that both methods assume independence

of the features (Molnar, 2022), which could have affected the obtained results. Conclusions drawn from SHAP values and partial dependence plots, however, could aid in pre-screening wells where modeling could be applicable.

For the parameters of the IRF, moderate correlations with the independent topographic variables were observed besides both the geological and climatic variables. Among these, variables describing the topographic location were present for the gain parameter *A* (such as the variable *hgthps* - the position of the well relative to the height range). A higher parameter *A* and *a* value was observed in the monitoring wells higher in the topography and with thicker vadose zones. Similarly, higher values of the parameter *n* which describes the time delay of the head response to a driver impulse, were observed in monitoring wells higher in the catchment (variable *arbel*). Some of the variables, however, can represent other processes and hence may be proxy values for different impacts. For example, the mean yearly temperature (*T_avg*) can also represent the orography, since lower temperatures on average are characteristic in the highlands (Bethere et al., 2017).

4.3. Limitations and possibilities

To investigate the correlations between the model parameters and the environmental variables, only 'good' models were selected. The criterion of model selection could be improved. In this study, a good fit corresponded to exceeding a metric threshold. Here, good is defined as passing a threshold of 0.65 for the NSE, which was arbitrarily set and a modelers' choice. If this threshold would be set to 0.5, already 818 and 380 models would have a good fit in the calibration and validation period, respectively (as opposed to 334 and 131 models, respectively, with the threshold of 0.65). Different selection criteria have been used in other studies, such as performing autocorrelation tests and determining the plausibility of the parameters (e.g., Brakenhoff et al., 2022; Zaadnoordijk et al., 2019). Within this study, the selection criterion impacts the regression and correlation analyses with the model parameters. The strength of correlations could change with different model selection criteria or with a different threshold value.

Over 40 environmental variables were used in the study to characterize the sites of the monitoring wells. The variables covered five groups (geologic, topographic, boundary, climatic, and land cover). However, the head is also influenced by other factors, such as soil texture (Chaudhuri and Ale, 2014; Szymkiewicz et al., 2019), but soil data is available at different quality and coverage in the Baltics (Kmoch et al., 2021). Furthermore, some of the variables were calculated over a specified radius (500 m). Since both local and regional processes influence head variability, multiple radii values could have been used to gain more insights. The influence of physical processes is not limited to a certain buffer distance, but rather process-based locations (for example, recharge and discharge points). To improve the representativeness of the environmental variables, a more thorough value calculation of the variables should be performed. There is further potential in making groundwater-related data more readily accessible for regular assessments and forecasts (Kmoch et al., 2016).

The dataset used for both correlation and regression analyses of the model fit (336 monitoring wells) was large enough for the insights to be representative. For the parameter studies, however, only 33 monitoring well data were used. Therefore, the results can present a distorted impression and should be interpreted carefully. Additionally, more environmental variables than time series were used, indicating that the random forest models could have been overfitted. A larger dataset for the parameter value studies could give more precise insight into the correlations. There are initiatives underway that could support accessibility to larger-scale harmonized data for such use cases that could be considered for groundwater data and the Baltics as well (Addor et al., 2017, 2020; Virro et al., 2021).

The SHAP values for multiple variables were very low indicating that only some of the variables impacted the predicted NSE value. Therefore, we chose to describe and analyze only the 5 most influential ones. However, the interpretation could be more insightful if the RF models were refitted and SHAP values recalculated with lower number of environmental variables. Thus, the regression could be improved by performing forward feature selection whereby a Random Forest model is fitted first on each variable separately. Models are further developed with the highest scoring variable and one of the others, again determining the highest score, and so on. It is repeated until adding an additional variable does not significantly improve the model fit (e.g., Virro et al., 2022; Macedo et al., 2019). Thus, the most influential variables are determined. Further, SHAP values could be subsequently analyzed (and mapped) for each separate monitoring well, which could give insights into regional differences of the most influencing environmental variables.

5. Conclusion

Head time series from 460 monitoring wells were modeled using time series models with impulse response functions. Only precipitation, potential evaporation, and temperature were used as driving forces. For each time series, four different model structures with different levels of complexity were used. From the 460 time series in the data set, 145 could be modeled well during the calibration period with at least one of the model structures. Only 68 time series could be modeled well with at least one of the models during the validation period. Of the four model structures, the nonlinear model with snow processes included (NLS) most often resulted in a good fit. This suggests that snow is an important process to take into account when modeling heads in the Baltics.

The fit of the model, measured as NSE in the calibration period, showed only weak correlations with the environmental variables. Some patterns, however, did appear from the results. The model fit correlates most strongly with the geological and climatic setting of the monitoring well. These results suggest that these models, in combination with the used driving forces, are most appropriate for modeling heads measured in monitoring wells with shallow groundwater and lower precipitation seasonality. Correlations between

parameter values and the environmental variables were also observed. Larger values for the gain and time lag parameters were generally found for locations with deeper water tables, less permeable sediments, and higher altitudes.

The relatively small number of time series that could be modeled well in this study may be explained by the limited number of driving forces that were used. We emphasize here that many of the wells are likely to be influenced by other driving forces such as groundwater pumping and surface water levels, and that adding these to the models may substantially improve the simulation of the heads. Nonetheless, we believe that the proposed methodology may help to gain a better understanding of the relationships between model performance, model parameters, and environmental and climatic settings. Both the good and the bad models provide insights that could direct future investigations into the groundwater in the Baltic states, for example as the initial analysis before the development of a distributed model or when input data is limited and a short calculation time is favored. To support future investigations using the same methodology, all the processed input data and code is made publicly available through a Zenodo repository (Jemeljanova and Collenteur, 2023).

Software implementation and data availability

The Sen's slopes were calculated with the *pyMannKendall* package (version 1.4.2.) (Hussain and Mahmud, 2019) for the *Python* programming language. The environmental variables were calculated using libraries *rgdal* (version 1.5–28), *rgeos* (version 0.5-9), *sp* (version 1.4-6), *sf* (version 1.0-8), and *Evapotranspiration* (version 1.16) in *R* programming language (R Core Team, 2021) (version 4.1.0). The head time series were modeled using the time series models as implemented in the *Python* open-source package *Pastas* (version 0.21.0) (Collenteur et al., 2019). The metrics were calculated and the Stoffer–Toloi test was performed using the *stats* sub-package from the *Pastas* package. The evapotranspiration was calculated using the PyET package (version 1.2.1.) (Matevž and Collenteur, 2021) for the *Python* programming language. *Scipy* package (version 1.6.3) for *Pyton* programming language was used for calculating the correlation coefficients, their statistical significance, and performing the Mann–Whitney U test. The *sklearn* package (version 1.1.3) was used for the Random Forest regression. The SHAP values were calculated using the TreeExplainer function from the *shap* package (version 0.41.0) (Lundberg and Lee, 2017) for *Python* programming language. The Partial Dependence was plotted using the *sklearn* package for *Python* programming language.

CRediT authorship contribution statement

Marta Jemeļjanova: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing. **Raoul A. Collenteur:** Conceptualization, Methodology, Writing – review & editing, Software. **Alexander Kmoch:** Conceptualization, Methodology, Writing – review & editing. **Jānis Bikše:** Conceptualization, Methodology, Formal analysis, Writing – review & editing. **Konrāds Popovs:** Formal analysis. **Andis Kalvāns:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data and scripts are available in a Zenodo repository (doi: http://dx.doi.org/10.5281/zenodo.7890699).

Acknowledgments

The authors would like to thank the Republic of Estonia Environment Agency, the Latvian Environment, Geology and Meteorology Centre, and the Lithuanian Geological Survey for providing the hydraulic head datasets for this research. We acknowledge the E-OBS dataset from the EU-FP6 project UERRA (https://www.uerra.eu) and the Copernicus Climate Change Service, and the data providers in the ECA&D project (https://www.ecad.eu). We thank the editor and the three anonymous reviewers for their valuable comments and suggestions.

Funding

The work of Marta Jemeļjanova, Jānis Bikše, and Konrāds Popovs was funded by the Latvian Council of Science, project "Spatial and temporal prediction of groundwater drought with mixed models for multilayer sedimentary basin under climate change", project No. lzp-2019/1-0165. The work of Andis Kalvāns was funded by the PostDoc research project agreement No. 1.1.1.2/VIAA/3/19/524. The work of Alexander Kmoch was funded by Estonian Research Agency grants No. PRG1764 and PSG841.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ejrh.2023.101416.

References

Addor, N., Do, H.X., Alvarez-Garreton, C., Coxon, G., Fowler, K., Mendoza, P.A., 2020. Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges. Hydrol. Sci. J. 65 (5), 712–725. http://dx.doi.org/10.1080/02626667.2019.1683182.

Addor, N., Newman, A.J., Mizukami, N., Clark, M.P., 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrol. Earth Syst. Sci. 21 (10), 5293–5313. http://dx.doi.org/10.5194/hess-21-5293-2017.

Amrhein, V., Greenland, S., McShane, B., 2019. Scientists rise up against statistical significance. Nature 567, http://dx.doi.org/10.1038/d41586-019-00857-9.

Babre, A., Kalvāns, A., Avotniece, Z., Retiķe, I., Bikše, J., Popovs, K., Jemeljanova, M., Zelenkevičs, A., Dēliņa, A., 2022. The use of predefined drought indices for the assessment of groundwater drought episodes in the Baltic States over the period 1989–2018. J. Hydrol.: Reg. Stud. 40, 101049. http://dx.doi.org/10.1016/j.ejrh.2022.101049.

Bakker, M., Maas, K., Schaars, F., von Asmuth, J.R., 2007. Analytic modeling of groundwater dynamics with an approximate impulse response function for areal recharge. Adv. Water Resour. 30, 493–504. http://dx.doi.org/10.1016/j.advwatres.2006.04.008.

Bakker, M., Maas, K., von Asmuth, J.R., 2008. Calibration of transient groundwater models using time series analysis and moment matching. Water Resour. Res. 44, W04420. http://dx.doi.org/10.1029/2007WR006239.

Bakker, M., Schaars, F., 2019. Solving groundwater flow problems with time series analysis: you may not even need another model. Groundwater 57, 826–833. http://dx.doi.org/10.1111/gwat.12927.

Baran, N., Surdyk, N., Auterives, C., 2021. Pesticides in groundwater at a national scale (France): Impact of regulations, molecular properties, uses, hydrogeology and climatic conditions. Sci. Total Environ. 791, 148137. http://dx.doi.org/10.1016/j.scitotenv.2021.148137.

Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A., Wood, E.F., 2018. Present and future Köppen-Geiger climate classification maps at 1-km resolution. Sci. Data 5, 180214. http://dx.doi.org/10.1038/sdata.2018.214.

Bennett, N.D., Croke, B.F., Guariso, G., Guillaume, J.H., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20. http://dx.doi.org/10.1016/j.envsoft.2012.09.011.

Bethere, L., Sennikovs, J., Bethers, U., 2017. Climate indices for the Baltic states from principal component analysis. Earth Syst. Dyn. 8, 951–962. http://dx.doi.org/10.5194/esd-8-951-2017.

Bierkens, M.F., Knotters, M., Geer, F.C.V., 1999. Calibration of transfer function-noise models to sparsely or irregularly observed time series. Water Resour. Res. 35 (6), 1741–1750. http://dx.doi.org/10.1029/1999WR900083.

Brakenhoff, D., Vonk, M., Collenteur, R., Baar, M., Bakker, M., 2022. Application of time series analysis to estimate drawdown from multiple well fields. Front. Earth Sci. 10, 907609. http://dx.doi.org/10.3389/feart.2022.907609.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Chaudhuri, S., Ale, S., 2014. Long-term (1930–2010) trends in groundwater levels in Texas: Influences of soils, landcover and water use. Sci. Total Environ. 490, 379–390. http://dx.doi.org/10.1016/j.scitotenv.2014.05.013.

Collenteur, R.A., 2022. Improved time series analysis of groundwater data through open-source software and better process representations. (Ph.D. thesis). University of Graz, Austria, URN: urn:nbn:at:at-ubg:1-181228.

Collenteur, R.A., Bakker, M., Caljé, R., Klop, S.A., Schaars, F., 2019. Pastas: Open source software for the analysis of groundwater time series. Groundwater 57, 877–885. http://dx.doi.org/10.1111/gwat.12925.

Collenteur, R.A., Bakker, M., Klammler, G., Birk, S., 2021. Estimation of groundwater recharge from groundwater levels using nonlinear transfer function noise models and comparison to lysimeter data. Hydrol. Earth Syst. Sci. 25, 2931–2949. http://dx.doi.org/10.5194/hess-25-2931-2021.

Cornes, R.C., van der Schrier, G., van den Besselaar, E.J., Jones, P.D., 2018. An ensemble version of the E-OBS temperature and precipitation data sets, J. Geophys. Res.: Atmospheres 123, 9391–9409, http://dx.doi.org/10.1029/2017JD028200.

Devia, G.K., Ganasri, B.P., Dwarakish, G.S., 2015. A review on hydrological models. Aquat. Proc. 4, 1001–1007. http://dx.doi.org/10.1016/j.aqpro.2015.02.126.

European Environment Agency, 2018. Corine Land Cover (CLC) 2018. https://land.copernicus.eu/pan-european/corine-land-cover/clc2018, Last accessed September 28, 2022.

Evans, C.D., Peacock, M., Baird, A.J., Artz, R.R.E., Burden, A., Callaghan, N., Chapman, P.J., Cooper, H.M., Coyle, M., Craig, E., Cumming, A., Dixon, S., Gauci, V., Grayson, R.P., Helfter, C., Heppell, C.M., Holden, J., Jones, D.L., Kaduk, J., Levy, P., Matthews, R., McNamara, N.P., Misselbrook, T., Oakley, S., Page, S.E., Rayment, M., Ridley, L.M., Stanley, K.M., Williamson, J.L., Worrall, F., Morrison, R., 2021. Overriding water table control on managed peatland greenhouse gas emissions. Nature 593, 548–552. http://dx.doi.org/10.1038/s41586-021-03523-1.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Ann. Statist. 29, http://dx.doi.org/10.1214/aos/1013203451.

Friedman, J.H., Meulman, J.J., 2003. Multiple additive regression trees with application in epidemiology. Stat. Med. 22 (9), 1365–1381. http://dx.doi.org/10.1002/sim.1501.

Geofabrik GmbH and Contributors of Open Street Map, 2018. OpenStreetMap data extracts. https://download.geofabrik.de/, Last accessed September 28, 2022.

Gupta, H.V., Kling, H., 2011. On typical range, sensitivity, and normalization of Mean Squared Error and Nash-Sutcliffe Efficiency type metrics. Water Resour. Res. 47, W10601. http://dx.doi.org/10.1029/2011WR010962.

Haaf, E., Giese, M., Heudorfer, B., Stahl, K., Barthel, R., 2020. Physiographic and climatic controls on regional groundwater dynamics. Water Resour. Res. 56, e2019WR026545. http://dx.doi.org/10.1029/2019WR026545.

Hargreaves, G.H., Samani, Z.A., 1985. Reference crop evapotranspiration from temperature. Appl. Eng. Agric. 1 (2), 96–99. http://dx.doi.org/10.13031/2013.26773.

Hendriks, D.M.D., Kuijper, M.J.M., van Ek, R., 2014. Groundwater impact on environmental flow needs of streams in sandy catchments in the Netherlands. Hydrol. Sci. J. 59, 562–577. http://dx.doi.org/10.1080/02626667.2014.892601.

Hocking, M., Kelly, B.F., 2016. Groundwater recharge and time lag measurement through Vertosols using impulse response functions. J. Hydrol. 535, 22–35. http://dx.doi.org/10.1016/j.jhydrol.2016.01.042.

Hussain, M., Mahmud, I., 2019. pyMannKendall: a python package for non parametric Mann Kendall family of trend tests. J. Open Source Softw. 4, 1556. http://dx.doi.org/10.21105/joss.01556.

Jaagus, J., Briede, A., Rimkus, E., Remm, K., 2010. Precipitation pattern in the Baltic countries under the influence of large-scale atmospheric circulation and local landscape factors. Int. J. Climatol. 30, 705–720. http://dx.doi.org/10.1002/joc.1929.

Jaagus, J., Sepp, M., Tamm, T., Järvet, A., Mõisja, K., 2017. Trends and regime shifts in climatic conditions and river runoff in Estonia during 1951–2015. Earth Syst. Dyn. 8, http://dx.doi.org/10.5194/esd-8-963-2017.

Jasechko, S., Birks, S.J., Gleeson, T., Wada, Y., Fawcett, P.J., Sharp, Z.D., McDonnell, J.J., Welker, J.M., 2014. The pronounced seasonality of global groundwater recharge. Water Resour. Res. 50, 8845–8867. http://dx.doi.org/10.1002/2014WR015809.

Jemeljanova, M., Collenteur, R.A., 2023. Modeling hydraulic heads with impulse response functions in different environmental settings (dataset). http://dx.doi.org/10.5281/zenodo.7890699. Last accessed: 08.05.2023.

Kalvāns, A., Popovs, K., Priede, A., Koit, O., Retiķe, I., Bikše, J., Dēliņa, A., Babre, A., 2021. Nitrate vulnerability of karst aquifers and associated groundwater-dependent ecosystems in the Baltic region. Environ. Earth Sci. 80, 628. http://dx.doi.org/10.1007/s12665-021-09918-7.

Kavetski, D., Kuczera, G., 2007. Model smoothing strategies to remove microscale discontinuities and spurious secondary optima im objective functions in hydrological calibration. Water Resour. Res. 43, W03411. http://dx.doi.org/10.1029/2006WR005195.

Kitterød, N.-O., Kværner, J., Aagaard, P., Arustienė, J., Bikše, J., Dagestad, A., Gundersen, P., Hansen, B., Hjartarson, A., Karro, E., Klavins, M., Marandi, A., Radienė, R., Retike, I., Rossi, P.M., Thorling, L., 2022. Hydrogeology and groundwater quality in the Nordic and Baltic countries. Hydrol. Res. 53, 958–982. http://dx.doi.org/10.2166/nh.2022.018.

Kløve, B., Ala-aho, P., Bertrand, G., Boukalova, Z., Ertürk, A., Goldscheider, N., Ilmonen, J., Karakaya, N., Kupfersberger, H., Kvœrner, J., Lundberg, A., Mileusnić, M., Moszczynska, A., Muotka, T., Preda, E., Rossi, P., Siergieiev, D., Šimek, J., Wachniew, P., Angheluta, V., Widerlund, A., 2011. Groundwater dependent ecosystems. Part I: Hydroecological status and trends. Environ. Sci. Policy 14, 770–781. http://dx.doi.org/10.1016/j.envsci.2011.04.002.

Kmoch, A., Kanal, A., Astover, A., Kull, A., Virro, H., Helm, A., Pärtel, M., Ostonen, I., Uuemaa, E., 2021. EstSoil-EH: a high-resolution eco-hydrological modelling parameters dataset for Estonia. Earth Syst. Sci. Data 13 (1), 83–97. http://dx.doi.org/10.5194/essd-13-83-2021.

Kmoch, A., Klug, H., Ritchie, A.B.H., Schmidt, J., White, P.A., 2016. A spatial data infrastructure approach for the characterization of New Zealand's groundwater systems. Trans. GIS 20 (4), 626–641. http://dx.doi.org/10.1111/tgis.12171.

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

Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F., 2006. World map of the Köppen-Geiger climate classification updated. Meteorol. Z. 15, 259–263. http://dx.doi.org/10.1127/0941-2948/2006/0130.

Kriauciuniene, J., Meilutyte-Barauskiene, D., Reihan, A., Koltsova, T., Lizuma, L., Sarauskiene, D., 2012. Variability in temperature, precipitation and river discharge in the Baltic States. Boreal Environ. Res. 17, ISSN: 1797-2469.

Latvian Geospatial information agency, 2022. Digital height model basic data. https://www.lgia.gov.lv/lv/atvertie-dati, Last accessed September 28, 2022.

Legates, D.R., McCabe, G.J., 1999. Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35 (1), 233–241. http://dx.doi.org/10.1029/1998WR900018.

Lipovetsky, S., Conklin, M., 2001. Analysis of regression in game theory approach. Appl. Stoch. Models Bus. Ind. 17, 319–330. http://dx.doi.org/10.1002/asmb.446.

Long, A.J., Mahler, B.J., 2013. Prediction, time variance, and classification of hydraulic response to recharge in two karst aquifers. Hydrol. Earth Syst. Sci. 17, 281–294. http://dx.doi.org/10.5194/hess-17-281-2013.

Lu, M., Rogiers, B., Beerten, K., Gedeon, M., Huysmans, M., 2021. Exploring river–aquifer interactions and hydrological system response using baseflow separation, impulse response modelling and time series analysis in three temperate lowland catchments. Hydrol. Earth Syst. Sci. Discuss. 26, 3629–3649. http://dx.doi.org/10.5194/hess-26-3629-2022.

Lukševičs, E., Stinkulis, G., Mūrnieks, A., Popovs, K., 2012. Geological evolution of the Baltic Artesian Basin. In: Dēliņa, A., Kalvāns, A., Saks, T., Bethers, U., Vircavs, V. (Eds.), Highlights of Groundwater Research in the Baltic Artesian Basin. University of Latvia, Riga, pp. 7–52.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2, 56–67. http://dx.doi.org/10.1038/s42256-019-0138-9.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates, Inc, pp. 4768–4777, ISBN: 9781510860964.

Macedo, F., Oliveira, M.R., Pacheco, A., Valadas, R., 2019. Theoretical foundations of forward feature selection methods based on mutual information. Neurocomputing 325, 67–89. http://dx.doi.org/10.1016/j.neucom.2018.09.077.

Mackay, J.D., Jackson, C.R., Wang, L., 2014. A lumped conceptual model to simulate groundwater level time-series. Environ. Model. Softw. 61, 229–245. http://dx.doi.org/10.1016/j.envsoft.2014.06.003.

Mander, Ü., Krasnova, A., Schindler, T., Megonigal, J.P., Escuer-Gatius, J., Espenberg, M., Machacova, K., Maddison, M., Parn, J., Ranniku, R., Pihlatie, M., Kasak, K., Niinemets, U., Soosaar, K., 2022. Long-term dynamics of soil, tree stem and ecosystem methane fluxes in a riparian forest. Sci. Total Environ. 809, 151723. http://dx.doi.org/10.1016/j.scitotenv.2021.151723.

Manzione, R.L., Soldera, B.C., Wendland, E.C., 2017. Groundwater system response at sites with different agricultural land uses: case of the Guarani Aquifer outcrop area, Brotas/SP-Brazil. Hydrol. Sci. J. 62, 28–35. http://dx.doi.org/10.1080/02626667.2016.1154148.

Marchant, B.P., Bloomfield, J.P., 2018. Spatio-temporal modelling of the status of groundwater droughts. J. Hydrol. 564, 397–423. http://dx.doi.org/10.1016/j.jhydrol.2018.07.009.

Marsala, R.Z., Capri, E., Russo, E., Bisagni, M., Colla, R., Lucini, L., Gallo, A., Suciu, N.A., 2020. First evaluation of pesticides occurrence in groundwater of Tidone Valley, an area with intensive viticulture. Sci. Total Environ. 736, 139730. http://dx.doi.org/10.1016/j.scitotenv.2020.139730.

Matevž, V., Collenteur, R., 2021. PyEt - a Python package to estimate potential and reference evapotranspiration. In: EGU General Assembly Conference Abstracts. pp. EGU21–15008. http://dx.doi.org/10.5194/egusphere-egu21-15008.

Molnar, C., 2022. Interpretable machine learning. A guide for making black box models explainable. URL https://christophm.github.io/interpretable-ml-book/.

Moriasi, D.N., Arnold, J.G., Liew, M.W.V., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE 50, 885–900. http://dx.doi.org/10.13031/2013.23153.

Moriasi, D.N., Gitau, M.W., Pai, N., Daggupati, P., 2015. Hydrologic and water quality models: Performance measures and evaluation criteria. Trans. ASABE 58, 1763–1785. http://dx.doi.org/10.13031/trans.58.10715.

Nachar, N., 2008. The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. Tutor. Quant. Methods Psychol. 4, http://dx.doi.org/10.20982/tqmp.04.1.p013.

Obergfell, C., Bakker, M., Zaadnoordijk, W.J., Maas, K., 2013. Deriving hydrogeological parameters through time series analysis of groundwater head fluctuations around well fields. Hydrogeol. J. 21, http://dx.doi.org/10.1007/s10040-013-0973-4.

Peterson, T.J., Western, A.W., 2014. Nonlinear time-series modeling of unconfined groundwater head. Water Resour. Res. 50, 8330–8355. http://dx.doi.org/10.1002/2013WR014800.

Peterson, T.J., Western, A.W., 2018. Statistical interpolation of groundwater hydrographs. Water Resour. Res. 54, 4663–4680. http://dx.doi.org/10.1029/2017WR021838.

Pogumirskis, M., Sīle, T., Seņņikovs, J., Bethers, U., 2021. PCA analysis of wind direction climate in the Baltic states. Tellus A 73, 1–16. http://dx.doi.org/10.1080/16000870.2021.1962490.

Popovs, K., Kalvāns, A., Jemeljanova, M., Saks, T., Dēliņa, A., Bikše, J., Babre, A., Retiķe, I., 2022. Bedrock surface topography of Latvia. J. Maps 1–12. http://dx.doi.org/10.1080/17445647.2022.2067011.

Popovs, K., Saks, T., Jātnieks, J., 2015. A comprehensive approach to the 3D geological modelling of sedimentary basins: example of Latvia, the central part of the Baltic Basin. Estonian J. Earth Sci. 64, 173–188. http://dx.doi.org/10.3176/earth.2015.25.

R Core Team, 2021. R: A language and environment for statistical computing. URL https://www.R-project.org/, Last accessed September 2, 2022.

Republic of Estonia Land Board, 2022. Elevation data 2017-2020. https://geoportaal.maaamet.ee/eng/Spatial-Data/Elevation-Data-p308.html, Last accessed September 28, 2022.

Retike, I., Bikše, J., Kalvāns, A., Dēliņa, A., Avotniece, Z., Zaadnoordijk, W.J., Jemeljanova, M., Popovs, K., Babre, A., Zelenkevičs, A., Baikovs, A., 2022. Rescue of groundwater level time series: How to visually identify and treat errors. J. Hydrol. 605, 127294. http://dx.doi.org/10.1016/j.jhydrol.2021.127294.

Sen, P.K., 1968. Estimates of the regression coefficient based on Kendall's Tau. J. Amer. Statist. Assoc. 63, 1379–1389. http://dx.doi.org/10.1080/01621459.1968.10480934.

Shapoori, V., Peterson, T.J., Western, A.W., Costelloe, J.F., 2015a. Estimating aquifer properties using groundwater hydrograph modelling. Hydrol. Process. 29, http://dx.doi.org/10.1002/hyp.10583.

Shapoori, V., Peterson, T.J., Western, A.W., Costelloe, J.F., 2015b. Top-down groundwater hydrograph time-series modeling for climate-pumping decomposition. Hydrogeol. J. 23, http://dx.doi.org/10.1007/s10040-014-1223-0.

Szymkiewicz, A., Savard, J., Jaworska-Szulc, B., 2019. Numerical analysis of recharge rates and contaminant travel time in layered unsaturated soils. Water (Switzerland) 11, 545. http://dx.doi.org/10.3390/w11030545.

Terasmaa, J., Retike, I., Vainu, M., Priede, A., Lode, E., Pajula, R., Koit, O., Tarros, S., Bikše, J., Popovs, K., 2020. Joint methodology for the identification and assessment of groundwater dependent terrestrial ecosystems in Estonia and Latvia. In: Water Resources Quality and Management in Baltic Sea Countries. Springer, pp. 253–275. http://dx.doi.org/10.1007/978-3-030-39701-2_12.

van Dijk, W.M., Densmore, A.L., Jackson, C.R., Mackay, J.D., Joshi, S.K., Sinha, R., Shekhar, S., Gupta, S., 2020. Spatial variation of groundwater response to multiple drivers in a depleting alluvial aquifer system, northwestern India. Progr. Phys. Geogr. 44, 94–119. http://dx.doi.org/10.1177/0309133319871941.

Virbulis, J., Bethers, U., Saks, T., Sennikovs, J., Timuhins, A., 2013. Hydrogeological model of the Baltic Artesian Basin. Hydrogeol. J. 21, 845–862. http://dx.doi.org/10.1007/s10040-013-0970-7.

Virro, H., Amatulli, G., Kmoch, A., Shen, L., Uuemaa, E., 2021. GRQA: Global river water quality archive. Earth Syst. Sci. Data 13 (12), 5483–5507. http://dx.doi.org/10.5194/essd-13-5483-2021.

Virro, H., Kmoch, A., Vainu, M., Uuemaa, E., 2022. Random forest-based modeling of stream nutrients at national level in a data-scarce region. Sci. Total Environ. 840, 156613. http://dx.doi.org/10.1016/j.scitotenv.2022.156613.

von Asmuth, J., Bierkens, M., 2005. Modeling irregularly spaced residual series as a continuous stochastic process. Water Resour. Res. 41 (12), http://dx.doi.org/10.1029/2004WR003726.

von Asmuth, J., Bierkens, M., Maas, K., 2002. Transfer function-noise modeling in continuous time using predefined impulse response functions. Water Resour. Res. 38 (12), 1287. http://dx.doi.org/10.1029/2001wr001136.

von Asmuth, J., Maas, K., Bakker, M., Petersen, J., 2008. Modeling time series of ground water head fluctuations subjected to multiple stresses. Ground Water 46 (1), 30–40. http://dx.doi.org/10.1111/j.1745-6584.2007.00382.x.

von Asmuth, J., Maas, K., et al., 2001. The method of impulse response moments: a new method integrating time series-, groundwater-and eco-hydrological modelling. In: Impact of Human Activity on Groundwater Dynamics. Proceedings of a Symposium Held During the Sixth IAHS Scientific Assembly, Maastricht, Netherlands, 18-27 July 2001. pp. 51–58.

Walsh, R.P.D., Lawler, D.M., 1981. Rainfall seasonality: Description, spatial patterns and change through time. Weather 36, 201–208. http://dx.doi.org/10.1002/j.1477-8696.1981.tb05400.x.

Wang, Y., Li, T., Hou, X., Zhang, Y., Li, P., 2022. Hydraulic modeling of water flow in the thick vadose zone under precipitation. Geoenviron. Disasters 9, 7. http://dx.doi.org/10.1186/s40677-022-00207-4.

Wunsch, A., Liesch, T., Broda, S., 2022. Deep learning shows declining groundwater levels in Germany until 2100 due to climate change. Nature Commun. 13, 1221. http://dx.doi.org/10.1038/s41467-022-28770-2.

Zaadnoordijk, W.J., Bus, S.A., Lourens, A., Berendrecht, W.L., 2019. Automated time series modeling for piezometers in the national database of the Netherlands. Groundwater 57, 834–843. http://dx.doi.org/10.1111/gwat.12819.