# Assessing automated gap imputation of regional scale groundwater level data sets with typical gap patterns

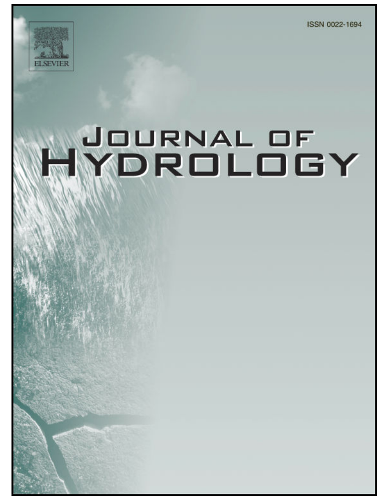


Citation for the original published paper (version of record):
Bikse, J., Retike, I., Haaf, E. et al (2023). Assessing automated gap imputation of regional scale groundwater level data sets with typical
gap patterns. Journal of Hydrology, 620. http://dx.doi.org/10.1016/j.jhydrol.2023.129424



(article starts on next page)

Journal Pre-proofs




Research papers

# Assessing automated gap imputation of regional scale groundwater level data sets with typical gap patterns

Jānis Bikše, Inga Retike, Ezra Haaf, Andis Kalvāns

PII: S0022-1694(23)00366-9
DOI: https://doi.org/10.1016/j.jhydrol.2023.129424
Reference: HYDROL 129424

To appear in: *Journal of Hydrology*

Received Date: 21 December 2022
Revised Date: 22 February 2023
Accepted Date: 17 March 2023

Please cite this article as: Bikše, J., Retike, I., Haaf, E., Kalvāns, A., Assessing automated gap imputation of regional scale groundwater level data sets with typical gap patterns, *Journal of Hydrology* (2023), doi: https://doi.org/10.1016/j.jhydrol.2023.129424

# Assessing automated gap imputation of regional scale groundwater level data sets with typical gap patterns

Jānis Bikše[1, *], Inga Retike[1], Ezra Haaf[2], Andis Kalvāns[1]

* Corresponding author: janis.bikse@lu.lv

[1] University of Latvia, Faculty of Geography and Earth Sciences,1 Jelgavas Street, LV-1004, Riga, Latvia.

[2] Chalmers University of Technology, Department of Architecture and Civil Engineering, SE-412 96 Gothenburg, Sweden.

## Abstract

Large groundwater level (GWL) data sets are often patchy with hydrographs containing continuous gaps and irregular measurement frequencies. However, most statistical time series analyses require regular observations, thus hydrographs with larger gaps are routinely excluded from further analysis despite the loss of coverage and representativity of an initially large data set. Missing values can be filled in with different imputation methods, yet the challenge is to assess the imputation performance of automated methods. Assessment of such methods tends to be carried out on randomly introduced missing values. However, large GWL data sets are commonly dominated by more complex patterns of missing values with longer contiguous gaps. This study presents a new artificial gap introduction approach (TGP-typical gap patterns) that improves our understanding of automated imputation performance by mimicking typical gap patterns found in regional scale groundwater hydrographs. Imputation performance of machine learning algorithm missForest and imputePCA is then compared with commonly applied linear interpolation to prepare a gapless daily GWL data set for the Baltic states (Estonia, Latvia, Lithuania). We observed that imputation performance varies among different gap patterns, and performance for all imputation algorithms declined when infilling previously unseen extremes and hydrographs influenced by groundwater abstraction. Further, missForest algorithm substantially outperformed other methods when infilling contiguous gaps (up to 2.5 years), while linear interpolation performs similarly for short random gaps. The TGP approach can be of use to assess the complexity of missing observation patterns in a data set and its value lies in assessing the performance of gap filling methods in a more realistic way. Thus the approach aids the appropriate selection of imputation methods, a task not limited to groundwater level time series alone. The study further provides insights into region-specific data peculiarities that can assist groundwater analysis and modelling.

Keywords: Time series; missing values; gap filling; the Baltic states; droughts; abstraction

# 1. Introduction

Assessment of groundwater resources at the regional scale is essential to sustainably manage transboundary aquifers and secure water supply (Kitterød et al., 2022; Wunsch et al., 2021). The EU Water Framework Directive requires member states to ensure good quantitative status of groundwater bodies by timely detection of negative trends posing a risk for resource depletion and deterioration of groundwater dependent ecosystems (WFD, 2000). Groundwater level time series or hydrographs are fundamental to evaluate the dynamics of a groundwater system (Zaadnoordijk et al., 2019). Often the groundwater level data sets compiled at the regional scales are patchy and spatially unevenly distributed (Barthel et al., 2021). Groundwater hydrographs rarely have equal observation periods and frequencies, and missing values are ubiquitous (Asgharinia and Petroselli, 2020; Peterson et al., 2017). The commonly required observation regularity is daily or weekly data (e.g. Haaf et al. (2020); Rajaee et al. (2019); Zanotti et al. (2019)), therefore, direct application of groundwater level time series in the further analysis is often hindered by the presence of gaps. Despite the omnipresent gaps and their negative impact on further usage of time series, there are no standardized methods for the estimation of missing groundwater heads (Dwivedi et al., 2022). Also, regional scale data sets are generally too large to infill gaps manually and individually. Therefore, it is more common to apply one automated method to all time series. The simplest approach is to remove any time series containing missing values (e.g. Nygren et al. (2020)) or those above a defined threshold (e.g. Heudorfer et al. (2019)). This approach is often justified by convenience, e.g., to fulfil the requirement of statistical software packages that do not support missing values (Dax and Zilberbrand, 2017; Emmanuel et al., 2021). Still, removing all time series with missing values (including gaps created during the correction process) may lead to a significant reduction of initially large groundwater level data sets. Retike et al. (2022) report that strict application of such an approach would lead to eight-fold reduction of number of available time series (from 612 to 76 groundwater hydrographs), while Haaf and Barthel (2018) had to remove almost 90% of data (from 4002 to 512 groundwater hydrographs). Wendt et al. (2020) removed all time series having more than 6 months long gaps and it reduced the initial data set from 660 to 170 hydrographs. To resolve this, imputation methods are used.

A common technique in handling missing observations in a time series is summary statistic imputation, that uses a single value – such as mean, mode or median obtained from the available observations to infill the missing values. However, this risks creating biases by introducing too many similar values (Pratama et al., 2016). Nevertheless, summary statistic imputation is routinely used to treat large databases because of its simplicity (e.g., Asgharinia and Petroselli (2020)). Even more common is linear interpolation for infilling relatively short gaps in groundwater level time series. The average length of linearly interpolated gaps in groundwater hydrographs by Lehr and Lischeid (2020) is 24.3 days (the maximum being 88 days). Likewise, Wunsch et al. (2021) interpolate up to one-month short sequences of missing data linearly. Sorensen et al. (2021) use linear interpolation to infill equal to or less than 3 months-long breaks given the slow response typically observed within the aquifers in drylands. While Wendt et al. (2020) apply linear interpolation to fill less than 6 months long data gaps. To sum up, simple imputation methods, like summary statistic imputation and linear interpolation from the last observation to the next, remain attractive to fill relatively short gaps in groundwater hydrographs.

The interest in using machine learning techniques to infill missing groundwater levels has increased due to their ability to deal with complex data sets (Dax and Zilberbrand, 2017; Dwivedi et al., 2022; Khedri et al., 2020; Rajaee et al., 2019; Vu et al., 2021). According to Retike et al. (2022) various gap patterns can be observed in national groundwater level databases, such as short gaps around extremes due to well completion problems or longer, contiguous gaps arising from malfunction or misplacement of automatic level loggers, and due to changes in monitoring programmes. For instance, Dwivedi et al. (2022) observe that it is relatively simple to accurately impute gaps in groundwater hydrographs that are introduced at random, but it is more challenging to infill contiguous gaps, especially around extremes. Deep learning techniques have been applied to impute as long as 47 years long gaps in groundwater hydrographs using piezometers as predictors that share the same hydrogeological context and characteristics with varying accuracy (RMSE 0.07 m to 1.08 m) (Vu et al., 2021). Oikonomou et al. (2018) emphasize that infilling missing values in groundwater hydrographs become an extremely difficult task when observations have low temporal and spatial frequency and represent highly dynamic groundwater systems. In such cases, most conventional data imputation methods are expected to fail. Approaches exist that reconstruct groundwater levels using observations from nearby wells (Dwivedi et al., 2022) or are supported with different types of auxiliary data, such as Earth observation data (e.g., Evans et al., 2020) or physiographic controls (Haaf et al., 2022). More advanced methods like missForest and imputePCA may be more promising and have hardly been tested on groundwater level data sets.

The non-parametric and iterative missForest method (Stekhoven and Bühlmann, 2012) is based on a random forest algorithm (Breiman, 2001) and designed for imputation of missing values by using the entire data set instead of imputation of one time series at a time. The approach has gained attention in several research fields including hydrology (Arriagada et al., 2021; Sidibe et al., 2018) and hydrogeology (Naranjo-Fernández et al., 2020). Among the major advantages of the missForest method are automatic and unsupervised missing data imputations that do not require assumptions about data distribution nor need tuning parameters. Further, the method is suitable for multivariate data sets with lots of missing data (Stekhoven and Bühlmann, 2012). Multiple authors report that missForest outperforms some well-known imputation methods, such as mean, nearest neighbour, linear regression and parametric methods (Alsaber et al., 2021; Arriagada et al., 2021; Stekhoven and Bühlmann, 2012; Waljee et al., 2013). The computational and thus time demand of the missForest, however, is large (Feng et al., 2014; Tang and Ishwaran, 2017). Another multiple imputation method is imputePCA (Josse and Husson, 2016), initially designed to impute missing data to perform principal components methods on incomplete data sets. The imputePCA has been used in meteorology (Benahmed and Houichi, 2018), but so far has not been applied in hydrogeology.

The performance of imputation algorithms can be assessed by introducing artificial missing values and then comparing the results of imputed values versus the original measurements. Several approaches exist how to choose lengths and distribution for artificial gaps in time series (Garciarena and Santana, 2017; Junninen et al., 2004). Often gaps are introduced at random with certain missing data percentages that can vary from 10% up to 90% of missingness in the data set (Dwivedi et al., 2022; Stekhoven and Bühlmann, 2012; Yadav and Roychoudhury, 2018). For instance, Arriagada et al. (2021) set thresholds that somewhat mimic the distribution of missing values in the raw data set. Afterwards, the modelled values are compared to the original observations using various metrics e.g., root mean square error (RMSE) and Nash-Sutcliffe efficiency (NSE) (Brakkee et al., 2022; Koch et al., 2019; Moriasi et al., 2015; Tao et al., 2022; Wang et al., 2018; Wunsch et al., 2022).

To understand how imputation methods perform with more erratic and longer gaps we present a new approach that introduces gaps that mimic typical gap patterns (TGP) found in regional scale data sets. Then we compare the performance of three data imputation methods (missForest, imputePCA and linear interpolation) through the evaluation of artificially introduced gaps with the proposed TGP and commonly applied random gap patterns (RGP).

## 2. Materials and Methods

### 2.1. Study region

The Baltic states (Estonia, Latvia and Lithuania) are located in north-eastern Europe. The region is characterized by a humid temperate climate affected by the Baltic Sea to the west and the Eurasian landmass to the east. According to the Köppen-Geiger classification, areas closer to the Baltic Sea are attributed to the temperate oceanic climate (Cfb), while the areas inland experience a warm-summer humid continental climate (Dfb) (Kottek et al., 2006). The median annual average temperature ranges from 5.3 to 8.6°C, precipitation from 566 to 770 mm/year (1991-2020, E-OBS data set: (Cornes et al., 2018)). The elevation varies from slightly below the sea level up to 318 m in the uplands. The land surface topography is defined to a large extent by the repeated advances of the quaternary ice sheets and their meltwater streams and lakes (Kalm et al., 2011; Zelčs et al., 2011). The region is characterized by distinct seasonality and negative temperatures with snow accumulation in the cold season. During the springtime, the first groundwater recharge maximum is driven by the snowmelt water infiltration, while in the period from September to December the second recharge maximum results from increased precipitation and low evapotranspiration (Babre et al., 2022). Generally, groundwater recharge takes place in the upland areas and discharges towards the Baltic Sea (Virbulis et al., 2013).

The Baltic states are situated on the multi-layered sedimentary Baltic Artesian Basin (BAB) with the thickness of sedimentary cover ranging from about 200 m in the North up to 6000 m in the South. The BAB comprises layers of clastic, carbonatic and in places evaporite sedimentary rocks gently dipping from northeast to southwest. The BAB holds vast amounts of groundwater with distinct chemical composition. The main aquifer systems used for water supply are formed of weakly cemented terrigenous and carbonate sedimentary rocks that fill the whole BAB overlain by Quaternary deposits (mostly glacial, glaciolacustrine, glaciofluvial and marine sediments). The aquifers are mostly confined, while unconfined aquifers are found in Quaternary sediments. A comprehensive overview of the hydrogeological setting and groundwater quality in the study region can be found in Kitterød et al. (2022).

### 2.2. Data and overall workflow

The workflow of this study (see Figure 1) was driven by the motivation to create gapless daily groundwater level time series for future research needs, namely, the identification of features controlling groundwater dynamics in the Baltic states and groundwater responses (rechange, groundwater-surface water interaction) at the event scale (e.g. using indices by Heudorfer et al., 2019). Thus, an effort was made to save as many wells as possible and to retain good spatial coverage.

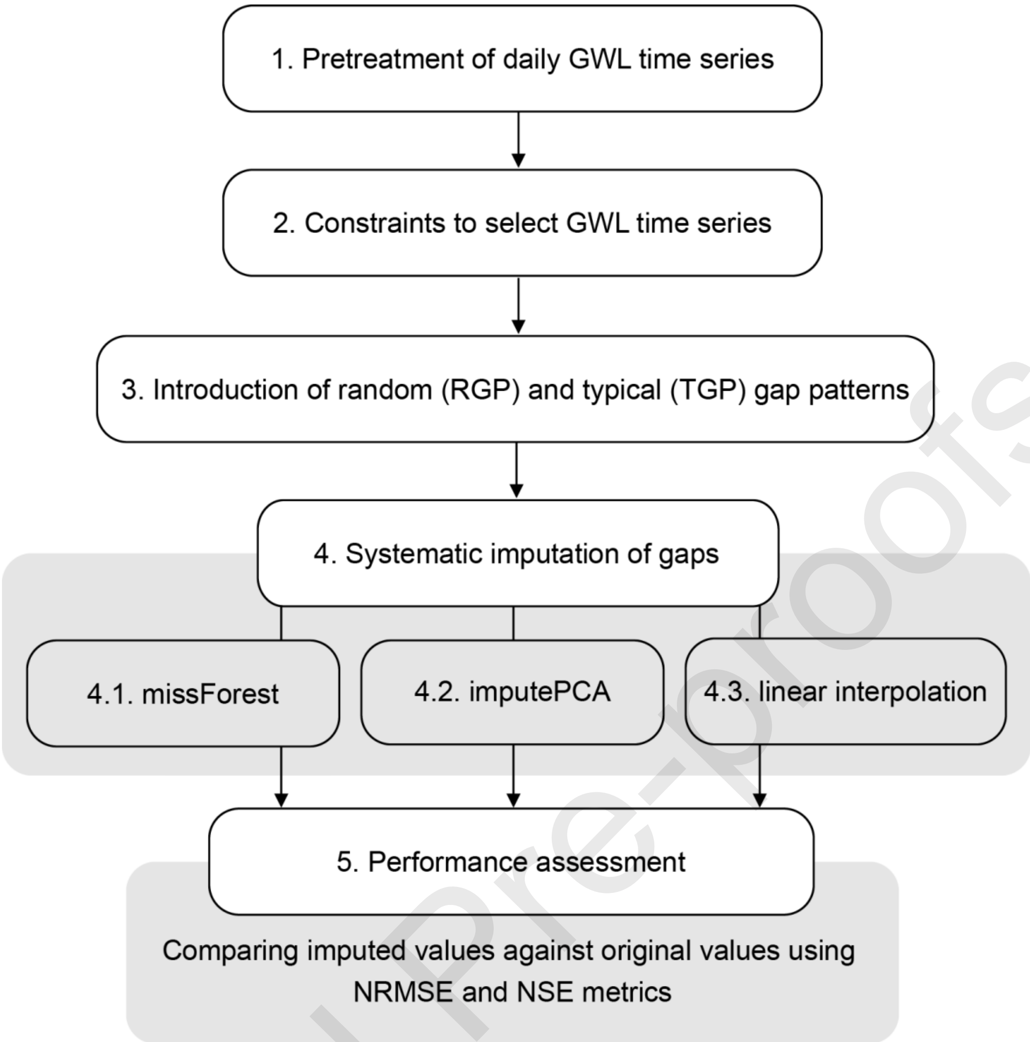



Figure 1. Workflow to create gapless groundwater level (GWL) time series and evaluate imputation performance (RGP – random gap patterns, TGP – typical gap patterns).

The raw groundwater level data were requested from the national managing authorities: The Estonian Environment Agency; Latvian Environment, Geology and Meteorology Centre; and Lithuanian Geological Survey. Groundwater level measurements from 465 wells with at least a single full month of daily observations were available for the period 2005 - 2021. The measurement frequency of the three data sets varied from daily (Lithuania) to sub-daily (Latvia, Estonia), where the latter was downsampled to daily frequency.

Combining the needs for adequate imputation performance assessment and future research, the final workflow contains following steps:

1. Pretreatment of daily groundwater hydrographs (step 1 in Figure 1) was carried out according to Retike et al. (2022) approach by visually screening for errors and applying necessary corrections. Pretreatment was applied to all 465 groundwater level time series. Out of total 2.66 million level measurements, 5.33% were deleted and 15.97% modified. Wells having anthropogenic influence (e.g., trends due to groundwater abstraction), as well as extreme drought episodes, were retained in the data set (Retike et al., 2022).
2. Selection of groundwater hydrographs according to the following constraints:

a. Temporal constraints. Identification of a period when most daily groundwater hydrographs are available for all three Baltic states - Estonia, Latvia, Lithuania.
b. Missingness and spatial constraints. This step involves identification of the maximum acceptable fraction of missing values in groundwater hydrographs that still can produce appropriate modelling results in combination with balancing the need for good spatial coverage of the wells.

3. The artificial gaps (step 3 in Figure 1) were introduced as missing values at random (random gap patterns - RGP) and as typical gap patterns (TGP).
4. Artificial gaps were infilled (step 4 in Figure 1) using missForest, imputePCA and linear interpolation methods in 12 calculation setups summarized in Table 1.
5. Imputation outcomes were assessed using NSE (Nash-Sutcliffe efficiency) and NRMSE (normalized root mean square error) metrics (step 5 in Figure 1) by evaluating relationships between the performance of applied models.

### 2.3. Preparation of artificial gaps

We clustered missing values in time series to identify the main gap patterns and their characteristics (such as length of consecutive gaps or spatial location of gaps (e.g. at the beginning, end or middle)) which could be further used to introduce artificial gaps. These groups mimic missing value patterns found in a particular data set. These groups, further called typical gap patterns (TGP) were identified in two steps. Firstly, groundwater hydrographs were transformed into a binary series by coding missing values as “0” and non-missing values as “1”. Then, on the binary series, TGP were identified using hierarchical cluster analysis (HCA) with a binary distance measure. The “agnes” function from R package “cluster” (version 2.1.2.) was used to carry out HCA with Ward linkage (Maechler et al., 2021; Murtagh and Legendre, 2014; Ward, 1963). Each cluster consisted of a number of hydrographs and if more than half of hydrographs had missing values in the same day, the whole day was treated as a missing value in a particular TGP. Random gap patterns were created from the data set with missingness rates of 10%, 20%, 30%, 40% and 50% using simple random sampling.

### 2.4. Imputation methods and procedure

#### 2.4.1. The missForest algorithm

The missForest algorithm (Stekhoven and Bühlmann, 2012) is based on a random forest (RF) algorithm where many decision trees are grown and averaged (Breiman, 2001). Simply speaking the algorithm uses donor hydrographs from the entire data set for target hydrograph imputation based on correlation of the non-missing parts between target and donors. Then, the gap in the target hydrograph is imputed through random forest regression based on all donor hydrographs.

Initially, missForest imputes all missing values with the target hydrographs mean value and then goes through all hydrographs with missing values (starting from the one with the least gaps). Then, a RF is built to predict the missing values. Each succeeding iteration builds a better model as previously imputed values are used as predictors for the next iteration. The imputation procedure is repeated several times until a stop criterion, or the pre-defined number of iterations is met.

In this study, the missForest was implemented using R version 4.2.0 (R Core Team, 2022) and the missForest package version 1.5 (Stekhoven, 2013) which permits to define parameters that are related to RF (such as ntree – number of trees to grow in each forest with 100 as a default number; and mtry - the number of variables randomly selected at each node to set up the split with the square of the number of variables as a default) and more specific missForest parameters (like maxiter– maximum number of iterations to be performed if stopping criteria is not yet met with a default value of 10). We tested multiple combinations of ntree (10 and 100) and maxiter (1 and 10) (see Table 1). To account for linear trends and seasons, new variables in addition to hydrographs were added (c.f. Evans et al., 2020) - day of the year, month and Unix time according to GWL observation date. The missForest imputation experiments were run on a different number of CPU cores by parallelizing on variables while preserving comparable computation times.

### 2.4.2. The imputePCA algorithm

The imputePCA algorithm is a statistical imputation method that imputes missing values by iteratively performing principal component analysis (PCA) on the data set. In this study variables are daily groundwater level time series including the gaps. The algorithm initially imputes missing observations with a mean value. Then PCA is carried out that finds the best approximation of the original hydrographs to reduce the initial number of dimensions which explain most of the variability within the data set. Next, a new prediction of the missing value is performed using the new dimensions and PCA analysis is run again. These steps are repeated until convergence, i.e. when the difference between two successive iterations is below a defined threshold or at a predefined number of iterations. Before imputation, the number of PCA dimensions must be estimated (Josse and Husson, 2016). Here, the missMDA R package was used to perform the imputePCA and 3 principal components were chosen for imputation based on the “estim_ncpPCA” output. The same additional variables (day of the year, month and Unix time) were added as in missForest setup (section 2.4.1.).

### 2.4.3. Linear interpolation

A commonly used linear interpolation method was applied to infill gaps in groundwater level times series. Linear interpolation imputed missing values by generating a straight line between two adjacent observations while missing values at the beginning or the end of a particular hydrograph were imputed by extrapolation of a constant value according to the first/last observation.

### 2.4.4. Imputation procedure

Each cluster representing TGP was used to introduce gaps in the original groundwater hydrographs. Then, missing values were imputed according to selected imputation methods. The missForest and imputePCA methods were applied for each groundwater hydrograph and each cluster separately, thus resulting in a relatively large number of imputations. Linear interpolation, however, was performed for each cluster for all hydrographs at once. The introduction of TGP gaps was carried out for all hydrographs, except for the time series, on which the gap pattern was created (see Section 2.2.), because these time series already had gaps in the same locations and could not be used to evaluate imputation performance. Moreover, groundwater hydrographs were excluded with more than 50% of missing values within the represented period of a particular cluster (TGP) to minimize the impacts of using too small data sets.

## 2.5. Model performance metrics

The performance of data imputation methods was evaluated by comparing the filled hydrographs with the observed data. This means that imputation was only evaluated for time series sections/values that originally did not contain gaps at the locations that were imputed. For evaluation Nash-Sutcliffe efficiency (NSE) and normalized root mean square error (NRMSE) were used. Metrics were calculated using the hydroGOF (Zambrano-Bigiarini, 2020) package in R.

# 3. Results

## 3.1. Selected groundwater level time series

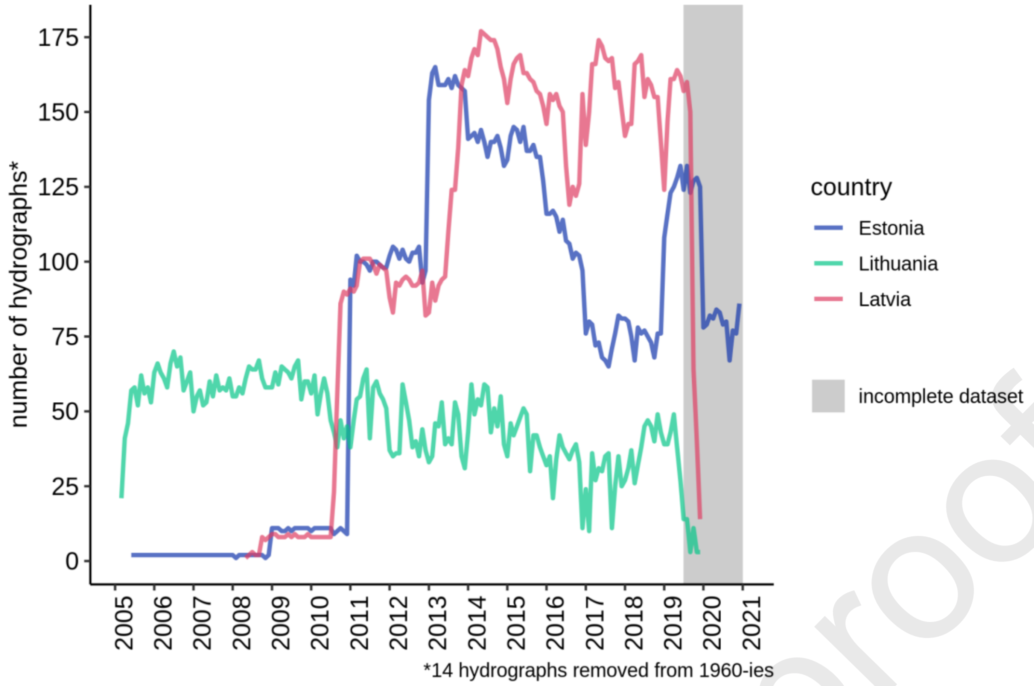



Figure 2. Changes in the number of daily hydrographs over time by countries (Data aggregated by months, only hydrograph sections having full month of daily observations shown).

The number of active groundwater level monitoring stations in the study area varied substantially between countries. In Lithuania, automatic-level loggers were deployed already in 2005 (Arustiene, 2011), while in Latvia (Retike et al., 2022) and Estonia the first automatic daily observations can be dated back to 2010-2011 (Figure 2). The number of available daily hydrographs for Lithuania remained relatively constant through the last two decades if compared to Latvia and Estonia where the changes were more dynamic (Figure 2). The peaks can be explained by the deployment of new automatic level loggers, while the sudden drops in observations most probably were due to various errors (mainly logger malfunction) and consequent data pretreatment, as well the exclusion of wells from monitoring programmes (Retike et al., 2022). It should be highlighted that abrupt drops at the end of the represented period (see grey shading in Figure 2) are associated with data gathering processes for this study and do not reflect the actual changes in the monitoring programmes.

We selected the period when daily hydrographs were available for all three countries simultaneously in a reasonable amount. The selected period spanned from 1st January 2011 to 31st May 2019 covering 3073 days (~8.4 years). Then the threshold of a minimum of 5.5 years of non-missing values was set for the selected period that resulted in 283 hydrographs out of the initial 465 time series. This criterion implies that up to 34.5% of the total length of time series were allowed to have missing values that could be either a continuous gap or a sum of several smaller gaps. Within the selected hydrographs consecutive lengths of missing values varied a lot – most hydrographs contained consecutive missing values with the maximum length up to about 100 days, while a significant part – 20% of hydrographs had maximum consecutive missing values of more than 1 year length (Figure S1 in Supplements). However, the average lengths of consecutive gaps within groundwater hydrographs were 55.2 days and a median was 27.8 days. Only 29 of the selected daily hydrographs had no missing values.

## 3.2. Introduction of artificial gaps

### 3.2.1. Typical gap patterns (TGP)

Eleven distinct groups of gaps in groundwater hydrographs were identified through cluster analysis of the binary missing value series in the data set (Figure 3a). Then, the clustering results were generalized into TGP to be used for the simulation of artificial gaps in time series (Figure 3b). After the generalization process, Cluster 1-10 (each formed by 9 to 38 hydrographs, see Figure 3a) had gaps, while cluster 11 was characterized as gapless. More detailed characteristics of each TGP can be found in the Supplementary material (Table S1).

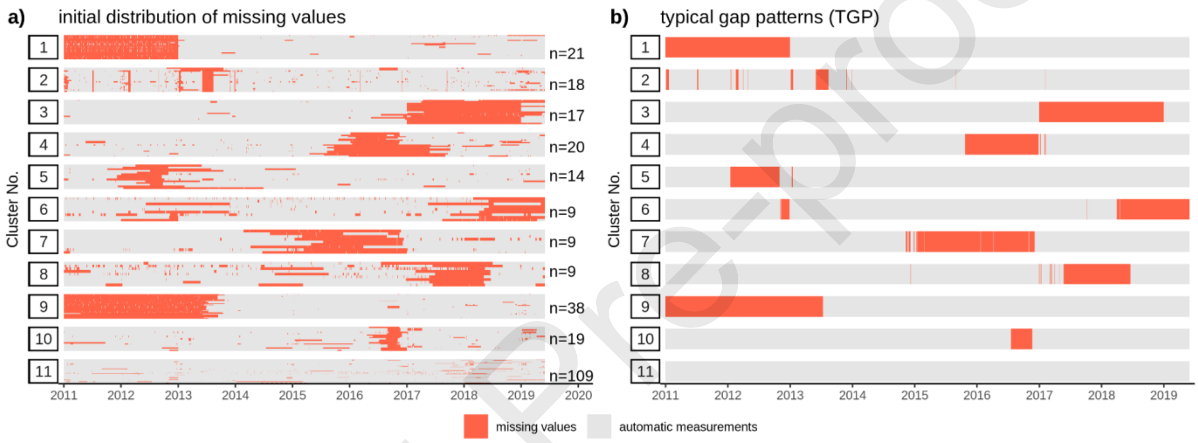



Figure 3. Patterns of missing values by clusters: a) initial distribution of missing values (n – number of hydrographs in the cluster); b) typical gap patterns (TGP) used to introduce artificial missing values in groundwater hydrographs. The height of the bars does not represent cluster size.

### 3.2.2. Random gap patterns (RGP)

The maximum cumulative length of gaps in the final data set (283 hydrographs) reached 1033 days or 33.6% of the total selected period, while on average hydrographs typically contained 14.1% of missing values (median value 11.29%). Therefore, RGP gaps were introduced only to the 109 hydrographs that form the gapless cluster 11 (Figure 3) using five thresholds of missingness (10%, 20%, 30%, 40% and 50%) based on the characteristics of the data set.

## 3.3. Imputation performance assessment

### 3.3.1. Overall imputation performance

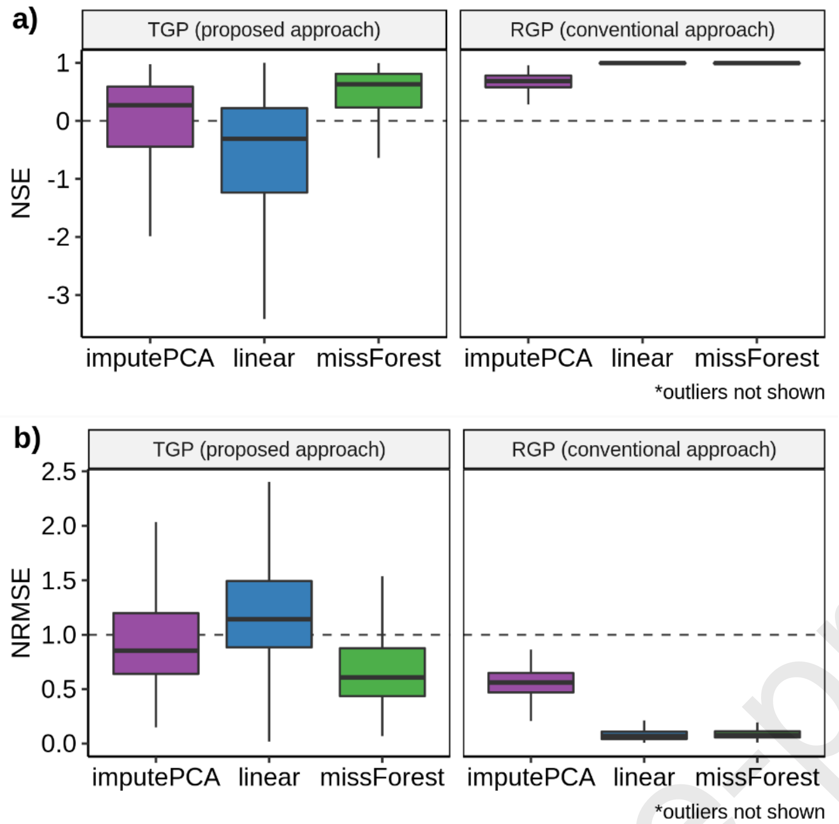



Figure 4. Overall performance of models on imputing typical gap patterns (TGP) and random gap patterns (RGP) as indicated by (a) NSE and (b) NRMSE values (Outliers not shown, see Supplementary Figure S2 with outliers and Table S2 for descriptive values).

In general, the imputation of RGP performed better than the imputation of TGP (Figure 4). Metrics showed similarly high performance for both, linear interpolation and missForest algorithms when infilling RGP (median NRMSE 0.07 and 0.08 respectively, median NSE 0.99 for both). In contrast, the performance of the imputePCA algorithm was less accurate and more dispersed (median NRMSE = 0.56, NSE = 0.69). The imputation of TGP was more challenging for all imputation algorithms and showed lower metric scores. The linear interpolation generally failed to impute missing values effectively (median NRMSE = 1.14, NSE = -0.31) and its performance sharply declined if compared to infilling RGP. On the contrary, the imputePCA algorithm demonstrated better performance compared to linear interpolation on TGP infilling (median NRMSE = 0.85, NSE = 0.27).

On average, the missForest algorithm outperformed the linear interpolation and imputePCA algorithms and regarding TGP showed the most satisfactory infilling results (median NRMSE = 0.61, NSE = 0.63). However, a proportion of hydrographs infilled by all algorithms, including missForest, showed extremely poor performance when imputing TGP, reaching as low NSE values as -1770, -170 and -14.5 and as high NRMSE as 3.9, 13.1 and 41.9 for imputePCA, missForest and linear interpolation, respectively.

### 3.3.2. Impact of individual gaps

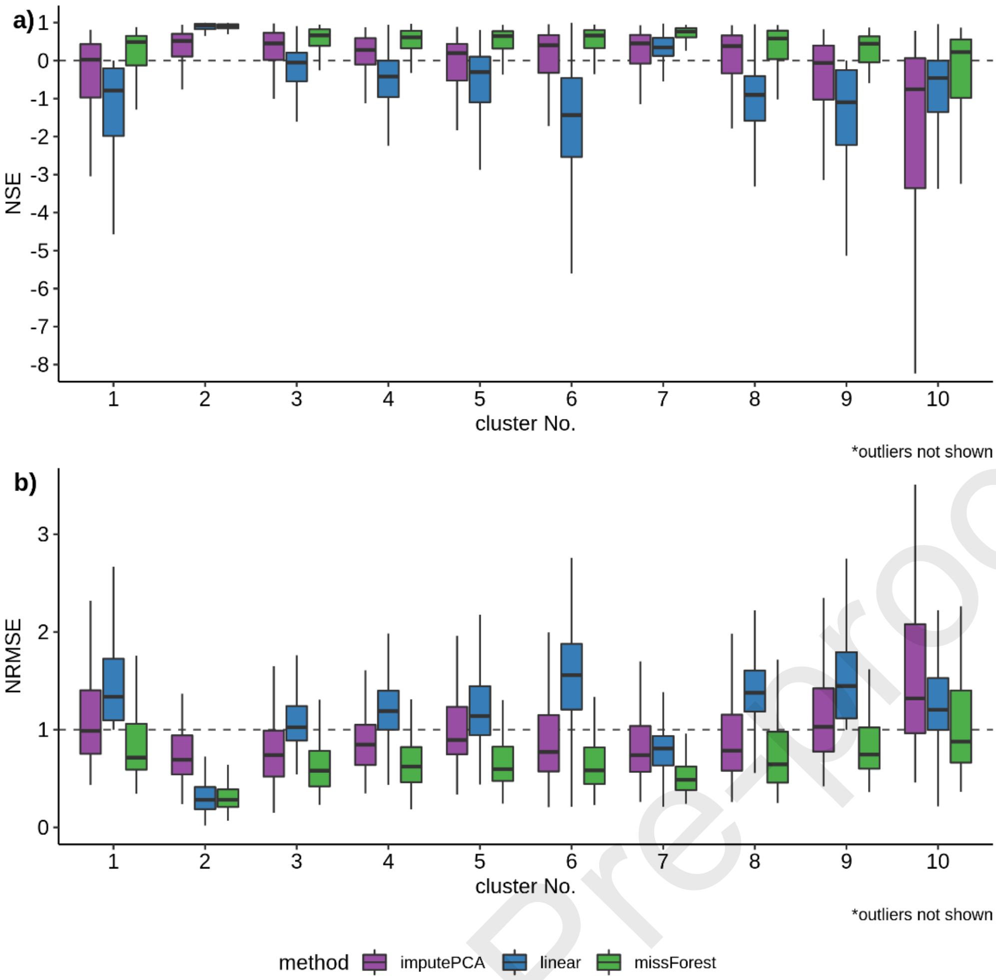



Figure 5. The performance of models on infilling each gap cluster of the TGP as indicated by (a) NSE and (b) NRMSE values (Dashed horizontal lines indicate thresholds for NSE and NRMSE; Outliers not shown, see Supplementary Figure S3 with outliers and table S2 for descriptive values).

The imputation performance of models on infilling TGP is relatively consistent (Figure 5). Overall, the missForest algorithm outperformed imputePCA and linear interpolation methods except for Cluster 2, where missForest and linear interpolation achieved similarly high imputation accuracy with median NRMSE below 0.28 and NSE above 0.92. In Cluster 2 there was less than 4.59% of missingness and the gaps were short, 7.83 days on average and thus more like the results of RGP.

It was observed that all methods performed poorly on infilling relatively long continuous gaps (2-2.5 years) located at the beginning of the hydrographs. For instance, in cluster 1 and 9, the missForest and imputePCA shows one of their worst results (median NRMSE >0.99 for imputePCA and >0.71 for missForest; median NSE for imputePCA <0.02 and missForest <0.49). The same low performance can be seen with linear interpolation, where gaps at the beginning or the end (Clusters 1, 6 and 9) are imputed with a constant first or last value in the hydrograph. In comparison, the accuracy of infilling similarly long contiguous gaps (2 years) not located at the ends of hydrographs (Cluster 3) was higher for missForest and imputePCA methods (median NRMSE for missForest 0.58, for imputePCA 0.74, median NSE for missForest 0.66, for imputePCA 0.45), with significantly lower scores for linear interpolation (NSE=-0.05, NRMSE=1.03). To sum up, the missForest algorithm performed well at infilling missing values of various complexity (short gaps, continuous gaps, and continuous gaps located in both ends of the time series). According to metrics the imputePCA often performed better than the linear interpolation (e.g., in infilling continuous gaps), but the variability of results was similarly high for both methods. Interestingly, that imputePCA performed substantially worse than linear interpolation in an apparently easy task such as infilling short gaps in Cluster 2.

### 3.3.3. Effect of missForest parameter setups

The four different parameter setups of missForest imputation we tested, significantly affected the computational times (see Table 1), while the impacts on the imputation performance were negligible (Figure 6). For example, TGP imputation of the computationally most demanding parameter setup (maxiter=10, ntree=100, the default settings of missForest) showed slightly improved performance compared to the weakest parameter setup (maxiter=1, ntree=10) (median NSE 0.63 and 0.56; median NRMSE 0.63 and 0.67), however, the spent computational time differs 37-fold (Table 1).

Reducing “ntree” from 100 to 10, decreased the required computational time 7 to 9-fold, while imputation performances are nearly the same (median NSE decrease: 0.06 and 0.003; median NRMSE rise: 0.05 and 0.02 for TGP and RGP respectively). Similar results were observed when the default “maxiter” was reduced from 10 to 1. The computational time decreased by 5 to 6-fold, while the impacts on the model performance were negligible (median NSE decreased by 0.01 and NRMSE increased by 0.01 when imputing TGP, but almost no changes in performance were observed when imputing RGP).

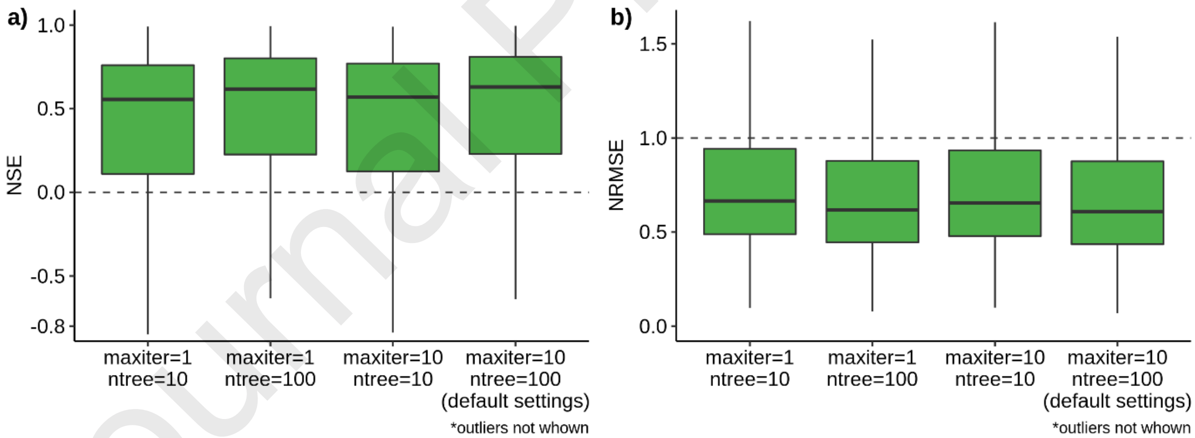



Figure 6. Effect of four different missForest parameter setups (maxiter – number of iterations; ntree – number of trees) on infilling TGP as indicated by (a) NSE and (b) NRMSE values (Outliers not shown, see Supplementary Figure S4 for outliers)

### 3.4. Imputation time analysis

For linear interpolation, imputation was performed for all hydrographs at once resulting in 10 iterations to impute all TGP (one for each TGP) and five iterations to impute the five RGP variants. The missForest and imputePCA methods required a single calculation for each TGP-hydrograph pair in a way that artificial missing values were introduced only on the selected hydrograph. As a result, 2384 unique TGP-hydrograph combinations were individually processed and imputed. RGP imputation for the two methods was performed in the same manner, but only on 109 hydrographs that form Cluster 11 resulting in 545 unique RGP-hydrograph combinations that were imputed for each model/experiment (Table 1).

The missForest imputation experiments were performed by using parallel processing natively implemented in the “doParallel" package (parallelizing on variables). ImputePCA experiments were parallelized by each TGP (10 steps) or by each RGP instance (5 steps). All imputations were performed on a high-performance computing server equipped with Intel Xeon Gold 5220R CPU. Imputation experiments were performed on a different number of CPUs, therefore computation times provided in Table 1 represent CPU time and not the actual time - i.e., the longest calculation with missForest (using parameters maxiter=10 and ntree=100) for TGP imputation (2384 individual imputations) took 392.47 CPU days, but because of running on 48 parallel CPUs, the actual spent time was only 8.18 days, that translates to ~5 minutes per single imputation.

Out of three imputation algorithms and tested experiments (Table 1), linear interpolation was the most rapid (10 to 22 seconds), followed by imputePCA (2.3 hours for RGP imputation and 92.1 hours for TGP imputation) and the missForest algorithm which had a wide range of computation times depending on the experimental setups (starting from 11.3 hours for RGP imputation using simple model parameters, up to 392.47 days for TGP imputation using default model parameters).

Table 1. Imputation experiment setups and imputation times. TGP-Typical gap patterns (2384 imputations); RGP-Random gap patterns (545 imputations)

| Model | Model parameters | Introduced gaps | Total CPU time [Days] | Average CPU time per imputation [minutes] | Total CPU time [seconds] | Number of CPUs used |
|---|---|---|---|---|---|---|
| missForest | maxiter=1, ntree=10 | TGP | 10.54 | 6.4 | 911024 | 16 |
| missForest | maxiter=1, ntree=100 | TGP | 71.84 | 43.4 | 6207136 | 16 |
| missForest | maxiter=10, ntree=10 | TGP | 49.72 | 30.0 | 4295472 | 16 |
| missForest | maxiter=10, ntree=100 | TGP | 392.47 | 237.1 | 33909312 | 48 |
| missForest | maxiter=1, ntree=10 | RGP | 0.47 | 1.2 | 40752 | 16 |
| missForest | maxiter=1, ntree=100 | RGP | 3.72 | 9.8 | 321440 | 16 |

| | | | | | | |
|---|---|---|---|---|---|---|
| missForest | maxiter=10, ntree=10 | RGP | 2.53 | 6.7 | 218912 | 32 |
| missForest | maxiter=10, ntree=100 | RGP | 22.57 | 59.6 | 1950112 | 32 |
| imputePCA | ncp=3 | TGP | 3.84 | 2.3 | 331560 | 10 |
| imputePCA | ncp=3 | RGP | 0.098 | 0.3 | 8430 | 5 |
| linear interpolation | - | TGP | 0.0003 | 0.0 | 22 | 1 |
| linear interpolation | - | RGP | 0.0001 | 0.0 | 10 | 1 |

## 4. Discussion

Despite the numerous techniques present on how to handle missing values (Junninen et al., 2004; Yadav and Roychoudhury, 2018), researchers tend to assess imputation performance by introducing artificial gaps at random (Dwivedi et al., 2022; Gill et al., 2007; Stekhoven and Bühlmann, 2012) as it is straightforward, fast and does not demand evaluation of the missingness patterns in the data set. However, our data set confirms the observation made in previous studies that missingness in time series data sets is dominated by salient gap patterns with larger gaps instead of random short gaps (Dwivedi et al. 2022; Oikonomou et al., 2018; Retike et al., 2022). From 11 TGP characteristic for the Baltic groundwater level data set, only one could be attributed to the random-like gaps (Cluster 2 with 18 on average 7.8 days long gaps). The present study also confirms that random-like gap patterns require less complex imputation methods to fill. Accordingly, the imputation of hydrographs from Cluster 2 achieved similarly good performance with linear interpolation and the missForest method (Figure 5). It is worth to highlight that groundwater level changes are often slow and consecutive changes between days can be small, thus it is easy to fill short gaps using the linear interpolation. Dwivedi et al. (2022) reported that even a simple interpolation method could accurately impute up to 90% of missing data in a two-year long period if gaps are at random, whereas the imputation of continuous gaps is much more challenging. This is in line with our findings as the linear interpolation method failed to impute the rest of the TGP adequately. In conclusion, assessing imputation methods based on TGP improves our understanding of imputation method performance.

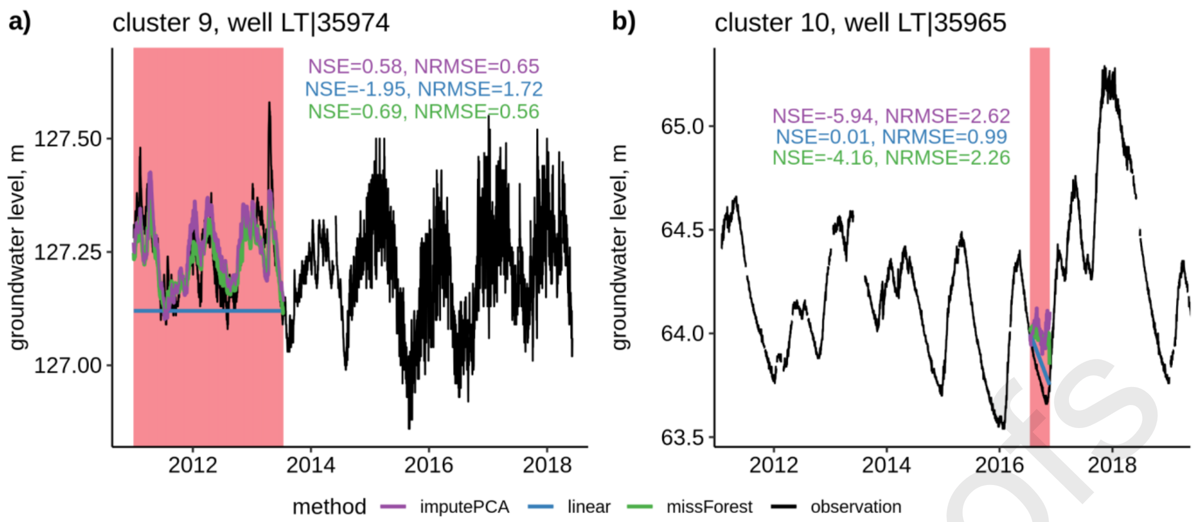



Figure 7. Imputed (red shaded area) hydrographs in (a) TGP Cluster 9 and (b) TGP cluster 10, showing that the length of a continuous gap is a poor predictor for imputation performance.

However, model performance for TGP is not solely related to gap length. Figure 7a shows a hydrograph with a long continuous gap (2.5 years) located at the challenging position - the beginning of the hydrograph, and still showing good imputation performance. While another hydrograph with a relatively shorter gap (126 days) located in the middle of the hydrograph showed lower performance (see Figure 7b). This could be due to the missingness of adequate (i.e., characteristic to the hydrograph) training data in a particular period since data-driven models learn the relationships solely from the input data Gill et al. (2007). However, at times the performance of models was unexpectedly poor even when infilling previously seen anomalies (such as in Figure 7b) and the true reason behind inaccurate infilling remains unclear. It should be remembered that the severity of gaps is a combination of how much data is missing, the mechanisms of missingness and the patterns of missing values (Emmanuel et al., 2021; Kang, 2013) not only the gap length itself.

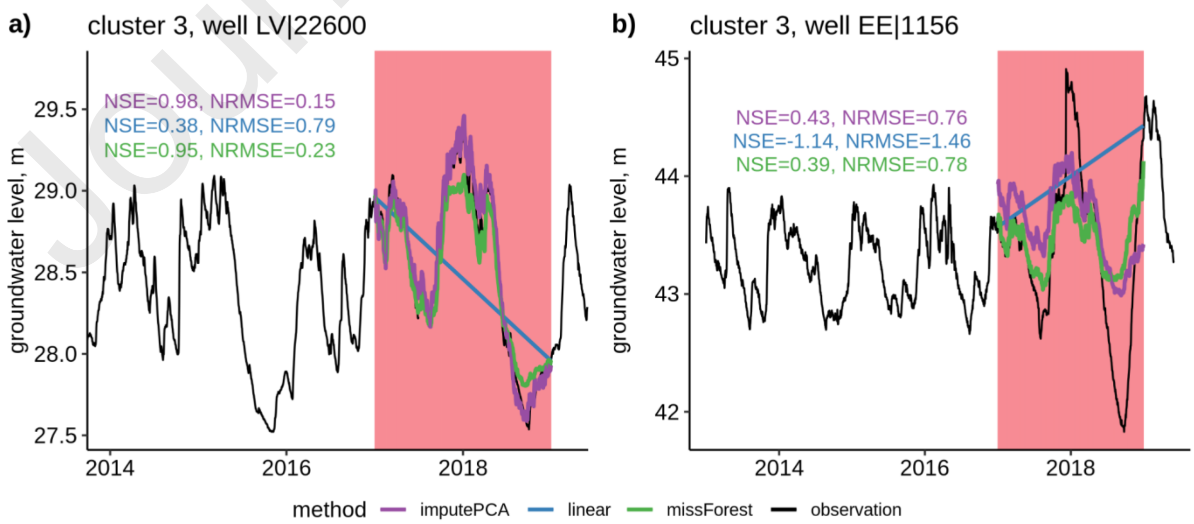


Figure 8. Imputed (red shaded area) hydrographs in TGP Cluster 3 showing relatively good (a) and bad (b) imputation performance most likely caused by previously seen (a) or unseen (b) anomalies.

When investigating method performance in individual time series, methods generally fail around gaps that contain extremes, as previously seen by e.g., Dwivedi et al. (2022). Figure 8 illustrates examples from Cluster 3, which contains a single continuous gap of 2 years from early 2017 to late 2018. The extremes are associated with a period of exceptionally severe drought episodes in Europe (Hänsel et al., 2022; Moravec et al., 2021; Rakovec et al., 2022), causing extreme groundwater droughts spanning to regional levels (Brakkee et al., 2022; Wunsch et al., 2022). While the performance of missForest and imputePCA was still satisfactory near peaks (Figure 8a), the performance of models declined in case of more severe (often previously unobserved) extremes (Figure 8b), probably as a result of untypical groundwater levels compared to the rest of the data set (Dwivedi et al., 2022). Therefore, we advise taking caution when imputing periods of extremes, independently of the imputation method.

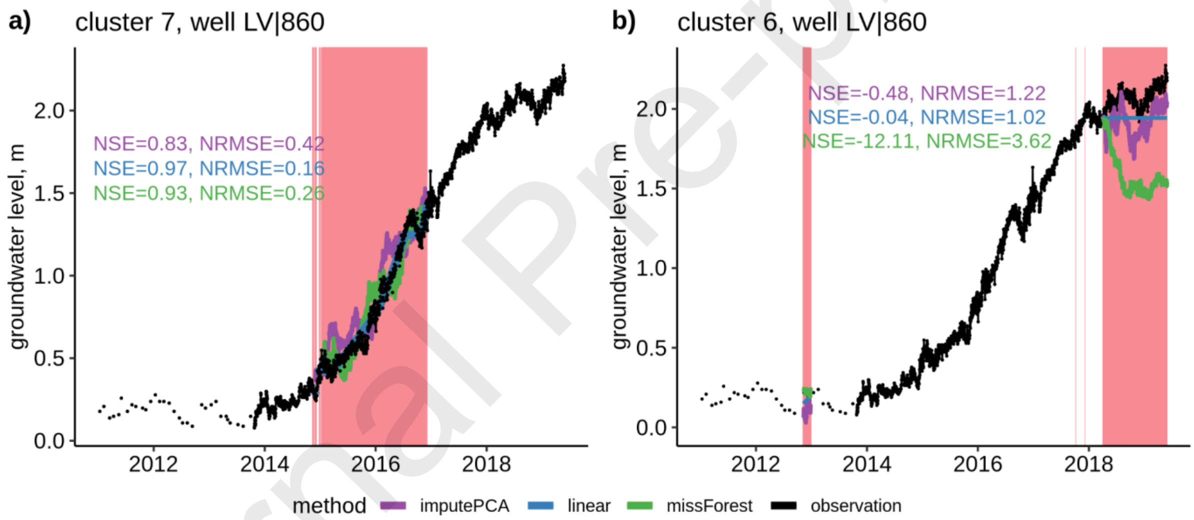



Figure 9. Imputed (red shaded area) gaps in TGP Cluster 7 (a) and Cluster 6 (b) in the same hydrograph, showing contrasting imputation performance due to anthropogenic impact and location of the gap.

In this study, the data set used for imputation did not include pre-selection constraints regarding anthropogenic influence on groundwater levels. However, intensive and continuous groundwater abstraction can cause deviations from expected groundwater level patterns affecting time series behaviour, such as trends or shifts (Brakkee et al., 2022; Sorensen et al., 2021). The groundwater hydrograph in Figure 9 is located in the vicinity of Liepāja (Latvia) historically known to be affected by groundwater over-abstraction (Bikše and Retike, 2018; Pulido-Velazquez et al., 2022) and illustrates the imputation performance in the presence of trends. It can be observed that linear interpolation can occasionally outperform more advanced imputation techniques and accurately impute longer gaps (maximum being ~1 year) in the hydrograph section with a rather stable trend (Figure 9a). At the same time, all three methods failed to handle shorter TGP at the end of the same hydrograph (Figure 9b), probably due to sudden changes in the time series pattern (Dwivedi et al., 2022).

The evaluation of abstraction effects on time series is often limited by the absence of actual water usage data (Wendt et al., 2020), including unregistered abstractions (Arriagada et al. 2021). Our study shows that poor imputation performance can be spatially linked to some documented cases of anthropogenically altered groundwater levels in Baltic states (Kitterød et al., 2022; Klimas et al., 2018) (Figure 10). Most negative NSE values are found near north-eastern Estonia, where extensive oil shale mining has significantly changed the groundwater levels (Terasmaa et al., 2020). Poor model performance can be observed for hydrographs near all three capital cities of the Baltic states (Tallinn, Riga and Vilnius) where groundwater is used as water supply (Kitterød et al. 2022; Klimas et al. 2018; Marandi and Karro 2008; Vallner and Porman 2016).

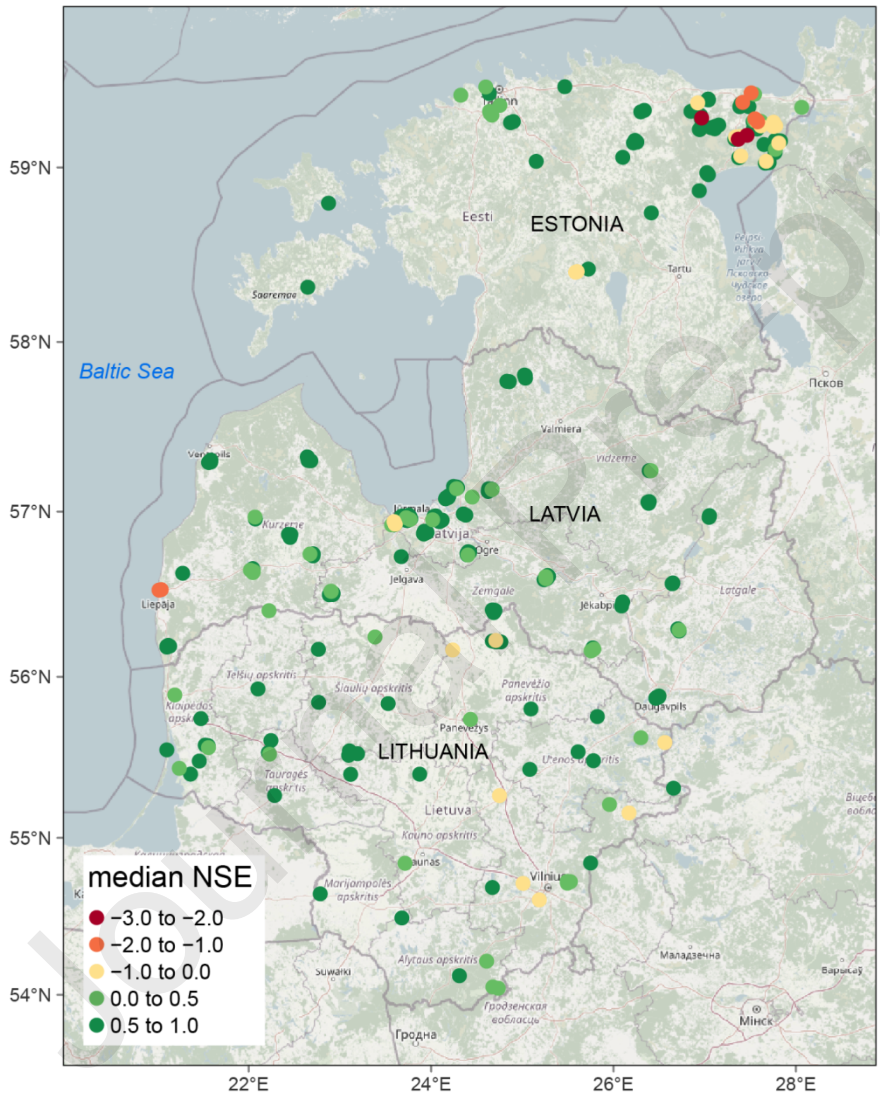



Figure 10. Spatial distribution of hydrographs and their median NSE values for missForest imputation of TGP (basemap from OpenStreetMap)

## 5. Conclusions

The present study demonstrates a new typical gap pattern (TGP) approach for assessing the performance of automated imputation methods considering the complexity of gap patterns in a data set, thus aiding the selection of the most appropriate imputation method. Most studies evaluating the infilling performance of various methods use random gaps, however, gaps found in groundwater level time series occur rarely at random. Therefore, an approach mimicking TGP found in daily groundwater hydrographs was developed. This approach was used to introduce artificial gaps and perform a more realistic in-depth evaluation of imputation method performance on patchy groundwater level data sets. Further, three data imputation methods (missForest, imputePCA and linear interpolation) were compared to infill introduced TGP and commonly applied random gaps on a set of regional scale groundwater hydrographs of the Baltic states. Overall, the missForest algorithm significantly outperformed both linear interpolation and imputePCA with TGP when infilling contiguous gaps (up to 2.5 years), often located near extremes and even at the beginning or end of the time series. However, with short, random gaps (7.8 days on average) the linear interpolation had as high imputation skill as missForest. The missForest default setup was the most time demanding and only slightly improved the performance, thus, could be recommended for single imputation needs. However, for TGP or similar multiple-imputation approaches, weaker parameter setups can be justified. We identified that in some cases low infilling performance arose due to previously unseen extremes such as the severe drought episode observed across Europe in 2018. Also, poor imputation performance could be attributed to known locations of intensive groundwater abstractions like the capital cities of the Baltic states – Tallinn, Riga and Vilnius. The proposed TGP approach affords imputation performance assessments higher granularity by systematically considering complex gap patterns and can be useful for large time series data sets beyond groundwater levels.

## Acknowledgements

This research is 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, Ministry of Education and Science, Latvia and University of Latvia grant No. ZD2016/AZ03 “Climate change and sustainable utilization of natural resources”. The research further contributes to grant TRV2019/45670 awarded by the Swedish Transport Administration (Trafikverket). The article was published with the financial support by the project No. 2018-1-0137 “EU-WATERRES: EU-integrated management system of cross-border groundwater resources and anthropogenic hazards” which benefits from a € 2.447.761 grant from Iceland, Liechtenstein and Norway through the EEA and Norway Grants Fund for Regional Cooperation.

## References

Alsaber, A.R., Pan, J., Al-Hurban, A., 2021. Handling Complex Missing Data Using Random Forest Approach for an Air Quality Monitoring Dataset: A Case Study of Kuwait Environmental Data (2012 to 2018). Int. J. Environ. Res. Public Health 18. https://doi.org/10.3390/ijerph18031333

Arriagada, P., Karelovic, B., Link, O., 2021. Automatic gap-filling of daily streamflow time series in data-scarce regions using a machine learning algorithm. J. Hydrol. 598. https://doi.org/10.1016/j.jhydrol.2021.126454

Arustiene, J., 2011. Groundwater Monitoring in Lithuania. In: Nalecz, T. (eds) Groundwater Management in the East of the European Union. NATO Science for Peace and Security Series C: Environmental Security. Springer, Dordrecht. https://doi.org/10.1007/978-90-481-9534-3_6

Asgharinia, S., Petroselli, A., 2020. A comparison of statistical methods for evaluating missing data of monitoring wells in the Kazeroun Plain, Fars Province, Iran. Groundwater for Sustainable Development 10, 100294. https://doi.org/10.1016/j.gsd.2019.100294

Babre, A., Kalvāns, A., Avotniece, Z., Retiķe, I., Bikše, J., Jemeljanova, K.P.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. Journal of Hydrology: Regional Studies 40, 101049. https://doi.org/10.1016/j.ejrh.2022.101049

Barthel, R., Haaf, E., Giese, M., Nygren, M., Heudorfer, B., Stahl, K., 2021. Similarity-based approaches in hydrogeology: proposal of a new concept for data-scarce groundwater resource characterization and prediction. Hydrogeology Journal 29, 1693-1709. https://doi.org/10.1007/s10040-021-02358-4

Benahmed, L., Houichi, L., 2018. The effect of simple imputations based on four variants of PCA methods on the quantiles of annual rainfall data. Environ. Monit. Assess. 190, 569. https://doi.org/10.1007/s10661-018-6913-y

Bikše, J., Retike, I., 2018. An Approach to Delineate Groundwater Bodies at Risk: Seawater Intrusion in Liepāja (Latvia). E3S Web of Conferences 54, 00003. https://doi.org/10.1051/e3sconf/20185400003

Brakkee, E., van Huijgevoort, M.H.J., Bartholomeus, R.P., 2022. Improved understanding of regional groundwater drought development through time series modelling: the 2018–2019 drought in the Netherlands. Hydrol. Earth Syst. Sci. 26, 551–569.

https://doi.org/10.5194/hess-26-551-2022

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Cornes, R.C., van der Schrier, G., van den Besselaar, E.J.M., Jones, P.D., 2018. An ensemble version of the E-OBS temperature and precipitation data sets. J. Geophys. Res. 123, 9391–9409. https://doi.org/10.1029/2017jd028200

Dax, A., Zilberbrand, M., 2017. Imputing missing groundwater observations. Hydrology Research 49, 831–845. https://doi.org/10.2166/nh.2017.220

Dwivedi, D., Mital, U., Faybishenko, B., Dafflon, B., Varadharajan, C., Agarwal, D., Williams, K.H., Steefel, C.I., Hubbard, S.S., 2022. Imputation of contiguous gaps and extremes of subhourly groundwater time series using random forests. JMLMC 3, 1-22. https://doi.org/10.1615/JMachLearnModelComput.2021038774

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., Tabona, O., 2021. A survey on missing data in machine learning. J Big Data 8, 140. https://doi.org/10.1186/s40537-021-00516-9

Evans, S., Williams, G. P., Jones, N. L., Ames, D. P., & Nelson, E. J., 2020. Exploiting Earth observation data to impute groundwater level measurements with an Extreme Learning Machine. Remote Sensing, 12(12), 2044. https://doi.org/10.3390/rs12122044

Water Framework Directive (WFD), 2000. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy.

Feng, L., Nowak, G., O'Neill, T.J., Welsh, A.H., 2014. CUTOFF: A spatio-temporal imputation method. J. Hydrol. 519, 3591–3605. https://doi.org/10.1016/j.jhydrol.2014.11.012

Garciarena, U., Santana, R., 2017. An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. Expert Syst. Appl. 89, 52–65. https://doi.org/10.1016/j.eswa.2017.07.026

Gill, M.K., Asefa, T., Kaheil, Y., McKee, M., 2007. Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique. Water

Resour. Res. 43, W07416. https://doi.org/10.1029/2006WR005298

Haaf, E., & Barthel, R., 2018. An inter-comparison of similarity-based methods for organisation and classification of groundwater hydrographs. Journal of Hydrology, 559, 222–237. https://doi.org/10.1016/j.jhydrol.2018.02.035

Haaf, E., Giese, M., Heudorfer, B., Stahl, K., Barthel, R., 2020. Physiographic and climatic controls on regional groundwater dynamics. Water Resour. Res. 56. https://doi.org/10.1029/2019wr026545

Haaf, Giese, Reimann and Barthel. 2022. Data-driven Estimation of Groundwater Level Time-Series Using Comparative Regional Analysis. *Authorea.* October 21, 2022. https://doi.org/10.1002/essoar.10512203.1

Hänsel, S., Hoy, A., Brendel, C., Maugeri, M., 2022. Record summers in Europe: Variations in drought and heavy precipitation during 1901–2018. Int. J. Climatol. 42, 6235–6257. https://doi.org/10.1002/joc.7587

Heudorfer, B., Haaf, E., Stahl, K., Barthel, R., 2019. Index-based characterization and quantification of groundwater dynamics. Water Resour. Res. 55, 5575–5592. https://doi.org/10.1029/2018wr024418

Josse, J., Husson, F., 2016. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. J. Stat. Softw. 70, 1–31. https://doi.org/10.18637/jss.v070.i01

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M., 2004. Methods for imputation of missing values in air quality data sets. Atmos. Environ. 38, 2895–2907. https://doi.org/10.1016/j.atmosenv.2004.02.026

Kalm, V., Raukas, A., Rattas, M., & Lasberg, K., 2011. Chapter 8 - Pleistocene Glaciations in Estonia. In J. Ehlers, P. L. Gibbard, & P. D. Hughes (Eds.), Developments in Quaternary Sciences (Vol. 15, pp. 95–104). Elsevier. https://doi.org/10.1016/B978-0-444-53447-7.00008-8

Kang, H., 2013. The prevention and handling of the missing data. Korean J. Anesthesiol. 64, 402–406. https://doi.org/10.4097/kjae.2013.64.5.402

Khedri, A., Kalantari, N., Vadiati, M., 2020. Comparison study of artificial intelligence method for short term groundwater level prediction in the northeast Gachsaran unconfined aquifer. Water Supply 20, 909–921. https://doi.org/10.2166/ws.2020.015

Kitterød, N.-O., Kværner, J., Aagaard, P., Arustienė, J., Bikše, J., Dagestad, A., Gundersen, P., Hansen, B., Hjartarson, Á., 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. Hydrology Research 53, 958–982. https://doi.org/10.2166/nh.2022.018

Klimas, A., Gregorauskas, M., Plankis, M., 2018. Management of the Wellfield Contaminated by Dense Chlorinated Hydrocarbons in Vilnius, Lithuania. Water Resour. 45, 238–249. https://doi.org/10.1134/S0097807818020094

Koch, J., Berger, H., Henriksen, H.J., Sonnenborg, T.O., 2019. Modelling of the shallow water table at high spatial resolution using random forests. Hydrol. Earth Syst. Sci. 23, 4603–4619. https://doi.org/10.5194/hess-23-4603-2019

Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F., 2006. World Map of the Köppen-Geiger climate classification updated. Meteorologische Zeitschrift. https://doi.org/10.1127/0941-2948/2006/0130

Lehr, C., Lischeid, G., 2020. Efficient screening of groundwater head monitoring data for anthropogenic effects and measurement errors. Hydrol. Earth Syst. Sci. 24, 501–513. https://doi.org/10.5194/hess-24-501-2020

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K., 2021. Cluster: Cluster Analysis Basics and Extensions. https://CRAN.R-project.org/package=cluster

Marandi, A., Karro, E., 2008. Natural background levels and threshold values of monitored parameters in the Cambrian-Vendian groundwater body, Estonia. Environ. Geol. 54, 1217–1225. https://doi.org/10.1007/s00254-007-0904-6

Moravec, V., Markonis, Y., Rakovec, O., Svoboda, M., Trnka, M., Kumar, R., Hanel, M., 2021. Europe under multi-year droughts: how severe was the 2014--2018 drought period? Environ. Res. Lett. 16, 034062. https://doi.org/10.1088/1748-9326/abe828

Moriasi, D.N., Gitau, M.W., Pai, N., Daggupati, P., 2015. Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria. Transactions of the ASABE 58, 1763–1785. https://doi.org/10.13031/trans.58.10715

Murtagh, F., Legendre, P., 2014. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J. Classification 31, 274–295. https://doi.org/10.1007/s00357-014-9161-z

Naranjo-Fernández, N., Guardiola-Albert, C., Aguilera, H., Serrano-Hidalgo, C., Montero-González, E., 2020. Clustering Groundwater Level Time Series of the Exploited Almonte-Marismas Aquifer in Southwest Spain. Water 12, 1063. https://doi.org/10.3390/w12041063

Nygren, M., Giese, M., Kløve, B., Haaf, E., Rossi, P.M., Barthel, R., 2020. Changes in seasonality of groundwater level fluctuations in a temperate-cold climate transition zone. Journal of Hydrology X 8, 100062. https://doi.org/10.1016/j.hydroa.2020.100062

Oikonomou, P.D., Alzraiee, A.H., Karavitis, C.A., Waskom, R.M., 2018. A novel framework for filling data gaps in groundwater level observations. Adv. Water Resour. 119, 111–124. https://doi.org/10.1016/j.advwatres.2018.06.008

Peterson, T.J., Western, A.W., Cheng, X., 2017. The good, the bad and the outliers: automated detection of errors and outliers from groundwater hydrographs. Hydrogeol. J. 26, 371–380. https://doi.org/10.1007/s10040-017-1660-7

Pratama, I., Permanasari, A.E., Ardiyanto, I., Indrayani, R., 2016. A review of missing values handling methods on time-series data, in: 2016 International Conference on Information Technology Systems and Innovation (ICITSI). pp. 1–6. https://doi.org/10.1109/ICITSI.2016.7858189

Pulido-Velazquez, D., Baena-Ruiz, L., Fernandes, J., Arnó, G., Hinsby, K., Voutchkova, D.D., Hansen, B., Retike, I., Bikše, J., Collados-Lara, A.J., Camps, V., Morel, I., Grima-Olmedo, J., Luque-Espinar, J.A., 2022. Assessment of chloride natural background levels by applying statistical approaches. Analyses of European coastal aquifers in different environments. Mar. Pollut. Bull. 174, 113303. https://doi.org/10.1016/j.marpolbul.2021.113303

Rajaee, T., Ebrahimi, H., Nourani, V., 2019. A review of the artificial intelligence methods in groundwater level modeling. J. Hydrol. 572, 336–351. https://doi.org/10.1016/j.jhydrol.2018.12.037

Rakovec, O., Samaniego, L., Hari, V., Markonis, Y., Moravec, V., Thober, S., Hanel, M., Kumar, R., 2022. The 2018–2020 multi-year drought sets a new benchmark in Europe.

Earths Future 10. https://doi.org/10.1029/2021ef002394

R Core Team, 2022. R: A Language and Environment for Statistical Computing.

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. https://doi.org/10.1016/j.jhydrol.2021.127294

Sidibe, M., Dieppois, B., Mahé, G., Paturel, J.-E., Amoussou, E., Anifowose, B., Lawler, D., 2018. Trend and variability in a new, reconstructed streamflow dataset for West and Central Africa, and climatic interactions, 1950–2005. J. Hydrol. 561, 478–493. https://doi.org/10.1016/j.jhydrol.2018.04.024

Sorensen, J.P.R., Davies, J., Ebrahim, G.Y., Lindle, J., Marchant, B.P., Ascott, M.J., Bloomfield, J.P., Cuthbert, M.O., Holland, M., Jensen, K.H., Shamsudduha, M., Villholth, K.G., MacDonald, A.M., Taylor, R.G., 2021. The influence of groundwater abstraction on interpreting climate controls and extreme recharge events from well hydrographs in semi-arid South Africa. Hydrogeol. J. 29, 2773–2787. https://doi.org/10.1007/s10040-021-02391-3

Stekhoven, D.J., 2013. missForest: Nonparametric Missing Value Imputation using Random Forest. R package version 1.4.

Stekhoven, D.J., Bühlmann, P., 2012. Missforest-Non-parametric missing value imputation for mixed-type data. Bioinformatics 28, 112–118. https://doi.org/10.1093/bioinformatics/btr597

Tang, F., Ishwaran, H., 2017. Random Forest Missing Data Algorithms. Stat. Anal. Data Min. 10, 363–377. https://doi.org/10.1002/sam.11348

Tao, H., Hameed, M.M., Marhoon, H.A., Zounemat-Kermani, M., Heddam, S., Kim, S., Sulaiman, S.O., Tan, M.L., Sa'adi, Z., Mehr, A.D., Allawi, M.F., Abba, S.I., Zain, J.M., Falah, M.W., Jamei, M., Bokde, N.D., Bayatvarkeshi, M., Al-Mukhtar, M., Bhagat, S.K., Tiyasha, T., Khedher, K.M., Al-Ansari, N., Shahid, S., Yaseen, Z.M., 2022. Groundwater level prediction using machine learning models: A comprehensive review. Neurocomputing 489, 271–308. https://doi.org/10.1016/j.neucom.2022.03.014

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: Negm, A.M., Zelenakova, M., Kubiak-Wójcicka, K. (Eds.), Water Resources Quality and Management in Baltic Sea Countries. Springer International Publishing, Cham, pp. 253–275. https://doi.org/10.1007/978-3-030-39701-2_12

Vallner, L., Porman, A., 2016. Groundwater flow and transport model of the Estonian Artesian Basin and its hydrological developments. Hydrology Research 47, 814–834. https://doi.org/10.2166/nh.2016.104

Virbulis, J., Bethers, U., Saks, T., Sennikovs, J., & Timuhins, A., 2013. Hydrogeological model of the Baltic Artesian basin. Hydrogeology Journal, 21(4), 845–862. https://doi.org/10.1007/s10040-013-0970-7

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

Waljee, A.K., Mukherjee, A., Singal, A.G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., Higgins, P.D., 2013. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open 3. https://doi.org/10.1136/bmjopen-2013-002847

Wang, X., Liu, T., Zheng, X., Peng, H., Xin, J., Zhang, B., 2018. Short-term prediction of groundwater level using improved random forest regression with a combination of random features. Applied Water Science 8, 125. https://doi.org/10.1007/s13201-018-0742-6

Ward, J.H., 1963. Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. 58, 236–244. https://doi.org/10.1080/01621459.1963.10500845

Wendt, D.E., Van Loon, A.F., Bloomfield, J.P., Hannah, D.M., 2020. Asymmetric impact of groundwater use on groundwater droughts. Hydrol. Earth Syst. Sci. 24, 4853–4868. https://doi.org/10.5194/hess-24-4853-2020

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

Wunsch, A., Liesch, T., Broda, S., 2021. Feature-based Groundwater Hydrograph Clustering

Using Unsupervised Self-Organizing Map-Ensembles. Water Resour. Manage. 36, 39–54. https://doi.org/10.1007/s11269-021-03006-y

Yadav, M.L., Roychoudhury, B., 2018. Handling missing values: A study of popular imputation packages in R. Knowledge-Based Systems 160, 104–118. https://doi.org/10.1016/j.knosys.2018.06.012

Zaadnoordijk, W.J., Bus, S.A.R., Lourens, A., Berendrecht, W.L., 2019. Automated Time Series Modeling for Piezometers in the National Database of the Netherlands. Ground Water 57, 834–843. https://doi.org/10.1111/gwat.12819

Zambrano-Bigiarini, M., 2020. hzambran/hydroGOF: v0.4-0. https://doi.org/10.5281/zenodo.3707013

Zanotti, C., Rotiroti, M., Sterlacchini, S., Cappellini, G., Fumagalli, L., Stefania, G.A., Nannucci, M.S., Leoni, B., Bonomi, T., 2019. Choosing between linear and nonlinear models and avoiding overfitting for short and long term groundwater level forecasting in a linear system. J. Hydrol. 578, 124015. https://doi.org/10.1016/j.jhydrol.2019.124015

Zelčs, V., Markots, A., Nartišs, M., & Saks, T., 2011. Chapter 18 - Pleistocene Glaciations in Latvia. In J. Ehlers, P. L. Gibbard, & P. D. Hughes (Eds.), Developments in Quaternary Sciences (Vol. 15, pp. 221–229). Elsevier. https://doi.org/10.1016/B978-0-444-53447-7.00018-0

Highlights

- A new approach to assess imputation performance considering gap patterns
- The approach mimics gaps found in regional scale groundwater level data set
- The MissForest algorithm outperforms imputePCA and linear interpolation in infilling daily groundwater hydrographs
- MissForest accurately imputes continuous gaps at the beginning or end of hydrographs, and around peaks
- However, hydrographs having unseen extremes (droughts) or water abstraction impacts are challenging to impute

CRediT author contribution statement

**Jānis Bikše:** conceptualization, writing – original draft, writing – review & editing, methodology, data curation, modelling, visualization **Inga Retike:** conceptualization, writing – original draft, writing – review & editing **Ezra Haaf:** writing – review & editing **Andis Kalvāns:** writing – review & editing, project administration