**Supplementary material**

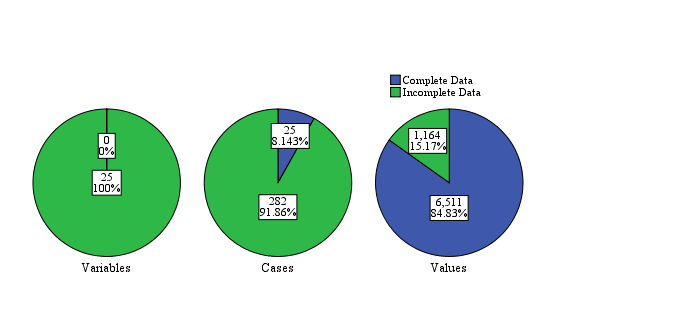
Reconstruction of groundwater level to impute missing values using singular and multichannel spectrum analysis: application to the Ardabil Plain, Iran

Majid Taie Semiromi and M. Koch

Corresponding author: orcid.org/0000-0003-4339-4217

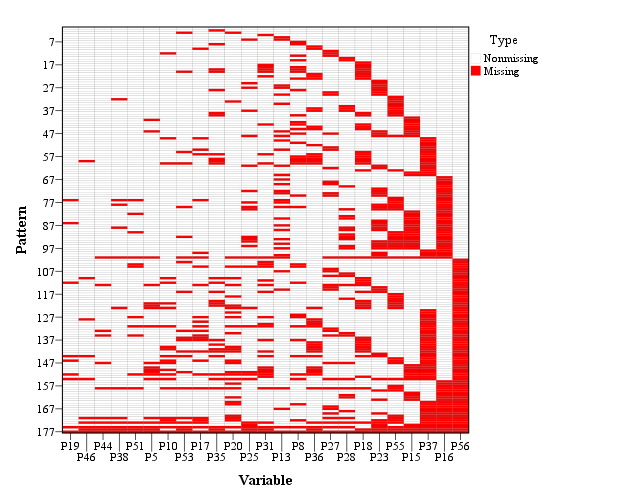
# 1 Statistical pattern characteristics of the observed data

To gain insight into the missing data of the groundwater level times series recorded in each piezometer, three pie charts were produced to show the overall missing values in piezometers (variables) and observations (cases), and a general pie chart of all data (values) (Fig. S1) (IBM 2011). The variables chart illustrates that each of the 25 piezometric stations contains at least one gap value on a case (observation). The cases chart indicates that 282 out of 307 observations has a minimum of one missing value for each piezometer. Finally, the values chart shows that 1164 of the theoretically total data, i.e. 6511 (307 × 25), which amounts to 15.17%, are missing.



**Figure S1.** Overall distribution of missing data in all 25 piezometric stations (variables) and the apportioned gaps throughout all observations (cases), as well as the percentage of missing and complete data (values).

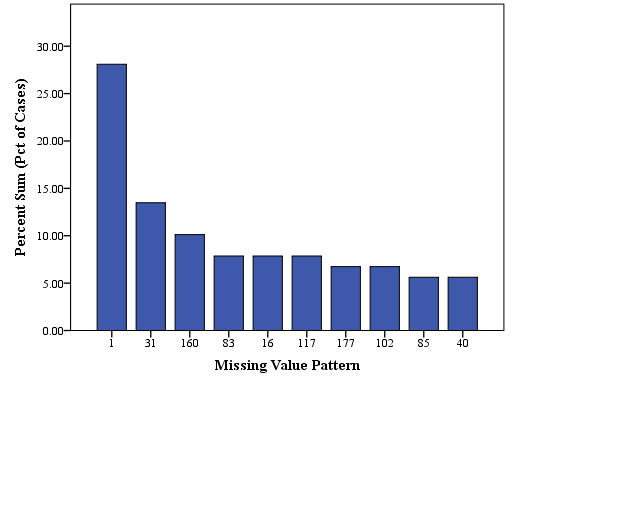
Also, we discerned that the way missing data arranged within the dataset has led to a certain pattern emerging – the arrangement of the hole (missing) locations of the piezometers when they are charted. As illustrated in Fig. S2, each pattern fits to a group of cases representing the same pattern of incomplete and complete data. For instance, Pattern 1 illustrates the cases that contain no missing values, whereas Pattern 3 shows cases that include missing values for the piezometric stations 20 and 53 (P20 and P53). A dataset can potentially contain two variable patterns (IBM 2011). In the current study, 177 patterns were detected within the 307 cases (times) for the dataset. In the case of a monotone pattern existing in the missing data, all missing and non-missing cells in the chart will be contiguous where no ‘islands’ of non-missing cells are seen in the lower right part of the chart and, similarly, no ‘islands’ of missing cells can be seen in the upper left portion of the chart (Little and Rubin 2002). Thus, following this missing and non-missing arrangement, the pattern of the missing values in the groundwater level time series recorded for 25 piezometric stations is found to be monotonic and, correspondingly, they need to be imputed.



**Figure S2.** Missing and non-missing patterns for 307 observations at the 25 piezometric stations.

Monotone missing data analysis has been taken into consideration through the literature, owing to reducing the mathematical complexity of maximum likelihood and multiple imputation and obviating the need for iterative estimation approaches (Enders 2010, Little and Rubin 2002).

Moreover, the frequency analysis performed for the 177 detected different patterns (Fig. S2) demonstrates that most observations fall into 10 pattern categories (Fig. S3). For instance, as illustrated, nearly 30% of the cases in the groundwater dataset show Pattern 1, and the missing value patterns chart (Fig. S2) indicates that this is the pattern for cases with no missing values (Fig. S2).



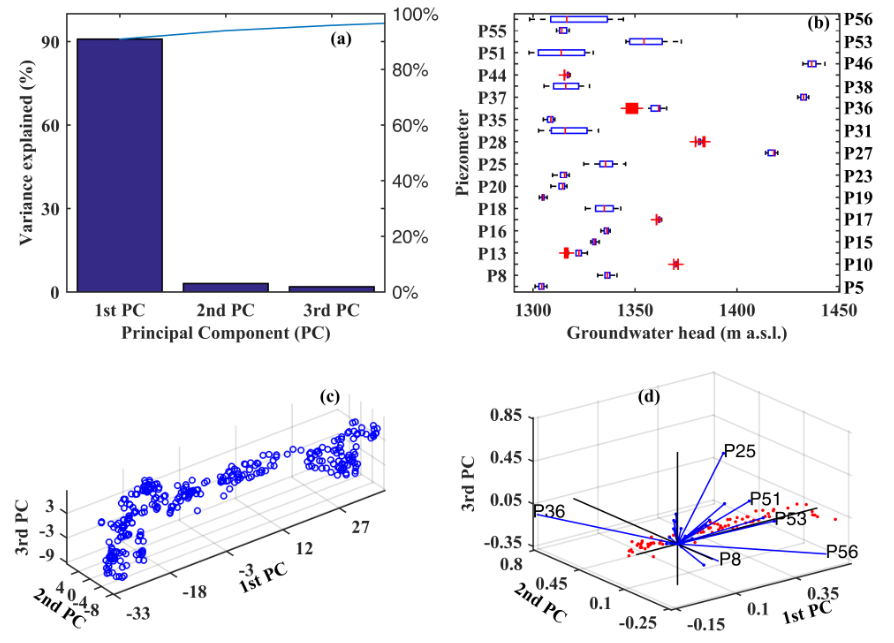
**Figure S3.** Frequency of 10 most recognized patterns (Fig. S2) extracted from the missing and complete data.

# 2. SSA objects

## 2.1. Principal component analysis (PCA)

As the SVD of the trajectory- or covariance-matrices is one of the major steps of SSA and MSSA (see equations (3) and (8) in the main article), the results of the SVD can also be employed directly for principal component analysis (PCA) of the possibly correlated multivariate set of the *p* = 25 piezometric time series. We forego a theoretical discussion of this widely used statistical tool and refer to e.g. (Jolliffe, 2002). Suffice to say that PCA consists in projecting the multivariate data in the direction of the axes (principal components) of a *p*-dimensional ellipsoid spanned up by the eigenvectors, with their lengths given by the associated singular or (singular) eigenvalues. By virtue of this fact, variances of the data in the directions of short axes (small singular values) are small and may be omitted from the dataset.

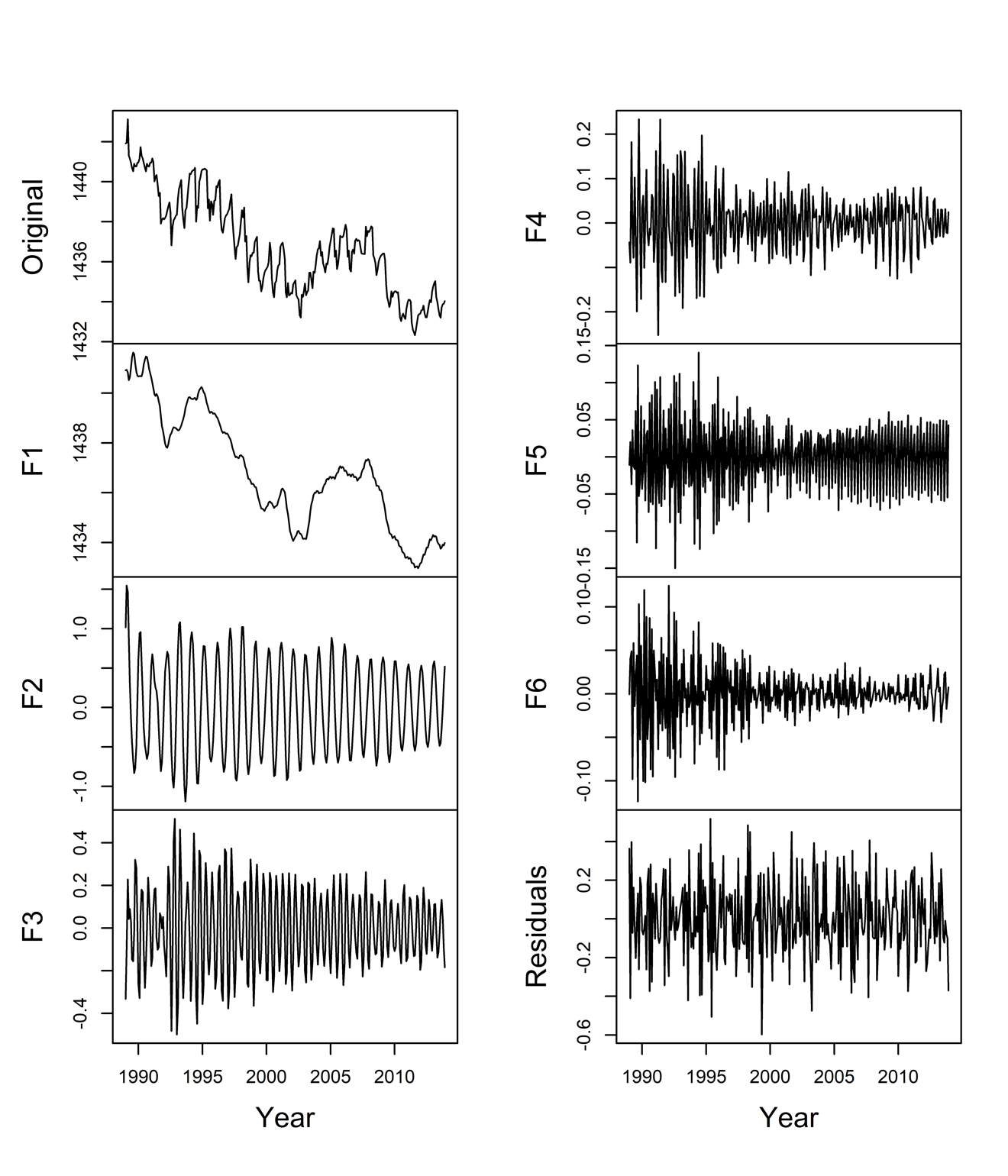
Figure S4 (a–d) illustrates some results of the PCA of the piezometric dataset. Figure S4(a) shows the variance explained by the first three principal components, indicating that the first component alone explains 90% of the total variance of the data. Figure S4(b) shows that the groundwater levels of piezometer P56 have the highest variability and for this reason they will be selected to illustrate SSA objects further down. Figure S4(c) shows how each observation for all piezometers is represented by these three major principal components. Finally, Figure S4(d) illustrates the contribution of six selected variables (piezometric series) to these three major principal components in terms of the corresponding vectors and, again, how each observation is represented by these components.



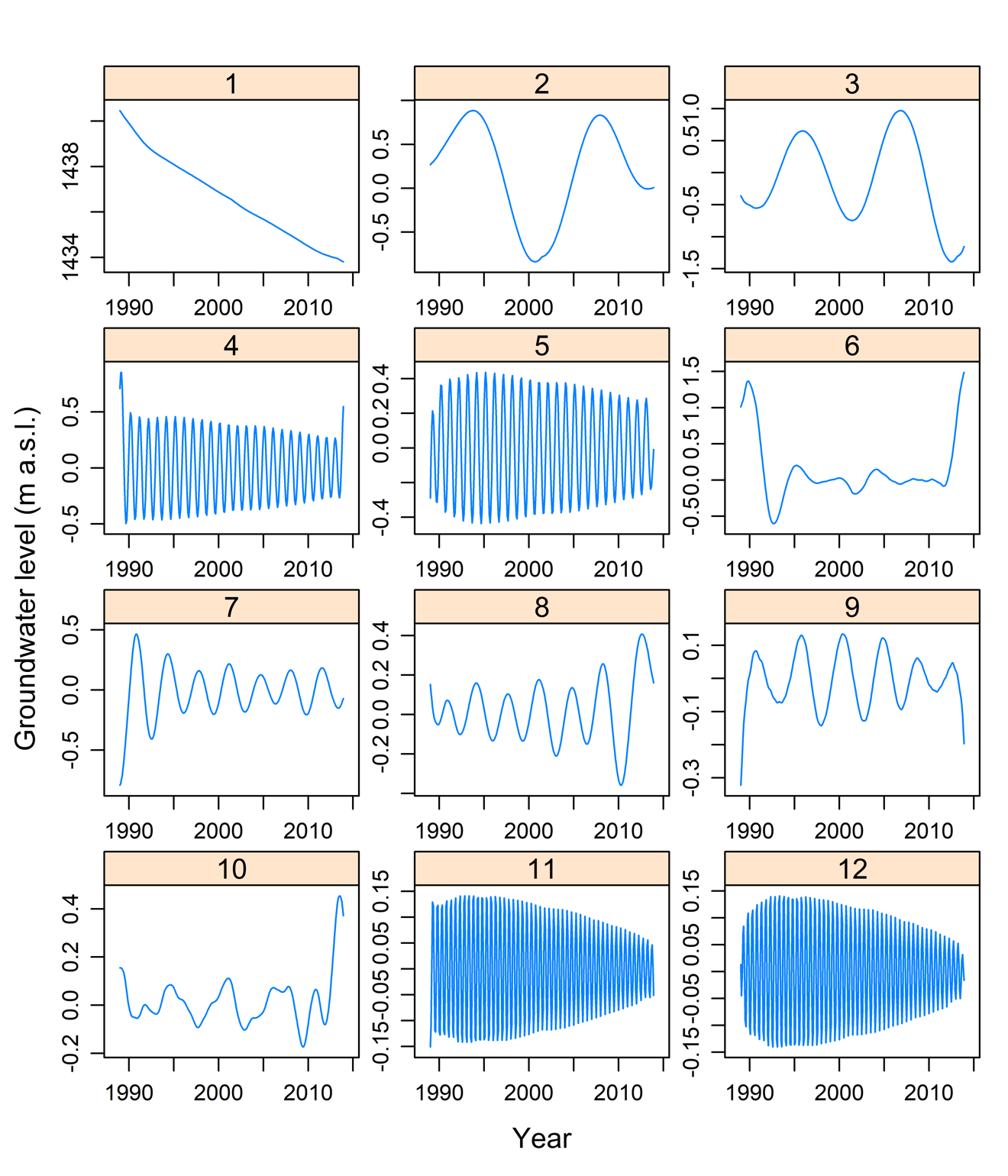
**Figure S4.** (a) Variance of the 25-variable multivariate piezometric dataset explained by the first three principal components. (b) Boxplots of the variability of the groundwater heads in the 25 piezometers. (c) Contribution of the observations to the major three principal components. (d) Biplots of the three principal components and the score matrix for six selected piezometers (variables), with the direction and length of the vectors indicating how each variable contributes to the three principal components in the plot and the points showing how each observation is represented by these three components.

## 2.2. SSA decomposition of piezometric time series

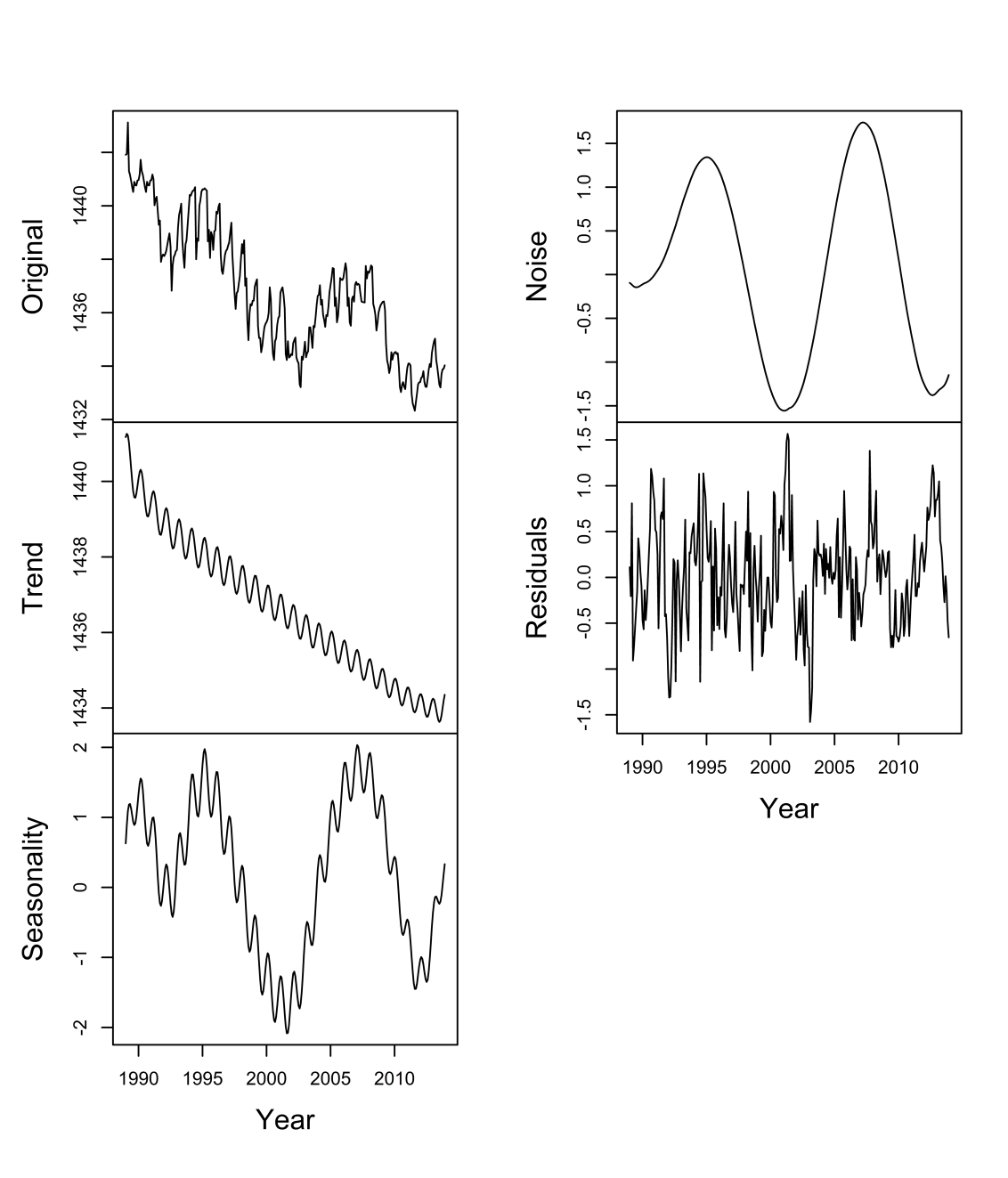
Figures S5–S7 show the results of the SSA-decomposition of the groundwater level series for piezometer P46 containing the minimum percentage of missing data (see Table 1) using a window length of *L* = 150, which is about half of the total length *N* of the series. In this regard, the groundwater level reconstructed by six eigentriples in the elementary grouping step and 12 eigenvectors in the final grouping process are illustrated by Figs S5 and S6, respectively. Both figures confirm that the first eigentriple and eigenvector correspond to the trend, while the other eigentriples contain high-frequency components and therefore are not related to the trend. The trend in Fig. S5 is closely reproduced by the first leading eigentriple and coincides with the first reconstructed component in Figs S5 and S6 for piezometric station P46. A full decomposition of groundwater time series measured in P46 is depicted by Fig. S7, as the time series is decomposed into trend, seasonality and noise. Due to a noticeable groundwater drawdown in this piezometer, a trend can be easily detected by the skill of these data driven models. However, the selection of window length affects the extent to which can be the trend extracted (Golyandina and Korobeynikov 2014). In addition, the groundwater fluctuations, representing mainly the recharge and discharge periods, are explained by detected seasonality. As indicated, SSA can satisfactorily distinguish signal and noise, sine waves with different frequencies, trend and seasonality (Golyandina and Zhigli︠a︡vskiĭ 2013).



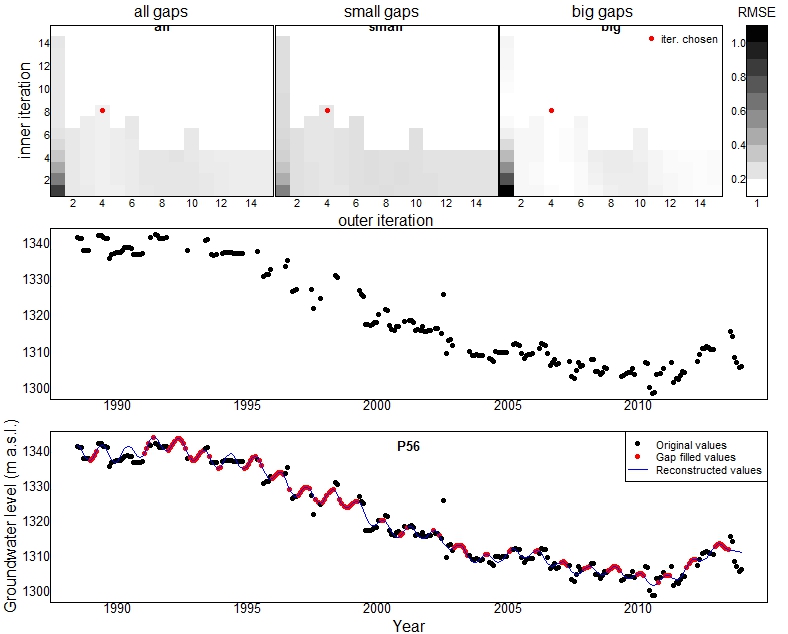
**Figure S5.** Initial reconstructed groundwater level for P46 using six eigentriples.



**Figure S6.** Reconstructed groundwater level in P46 using 12 leading eigenvectors.



**Figure S7.** Full decomposition of groundwater level time series of P46 into trend, seasonality and noise.



Small gaps

Outer iteration

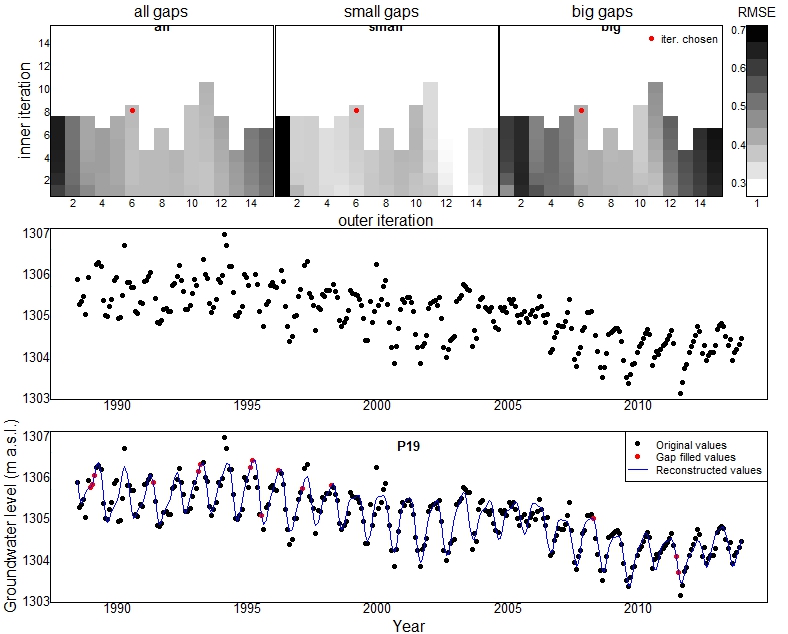
Inner iteration

Big gaps

All gaps

RMSE

**Figure S8.** Cross-validation analysis to identify optimal inner and outer iterations for P56.



Small gaps

All gaps

RMSE

Outer iteration

Inner iteration

Outer itiration

Big gaps

**Figure S9.** Cross-validation analysis to identify optimal inner and outer iterations for P19.

**References**

Enders, C. K., 2010. *Applied missing data analysis.* New York/London: Guilford.

Golyandina, N. and Korobeynikov, A. 2014. Basic Singular Spectrum Analysis and forecasting with R. *Computational Statistics & Data Analysis,* 71, 934-954.

Golyandina, N. and Zhigli︠a︡vskiĭ, A. A., 2013. *Singular spectrum analysis for time series.* Heidelberg: Springer.

IBM, 2011. *IBM SPSS Missing Values 20.* USA.

Little, R. J. A. and Rubin, D. B., 2002. *Statistical Analysis With Missing Data.* Second edition. Hoboken, NJ, USA: John Wiley & Sons.