Supplement of Hydrol. Earth Syst. Sci., 26, 551–569, 2022
https://doi.org/10.5194/hess-26-551-2022-supplement

Hydrology and Earth System Sciences
Open Access
EGU

Supplement of

Improved understanding of regional groundwater drought development through time series modelling: the 2018–2019 drought in the Netherlands

Esther Brakkee et al.

Correspondence to: Esther Brakkee (esther.brakkee@kwrwater.nl)

Supplement

S1 Example figures

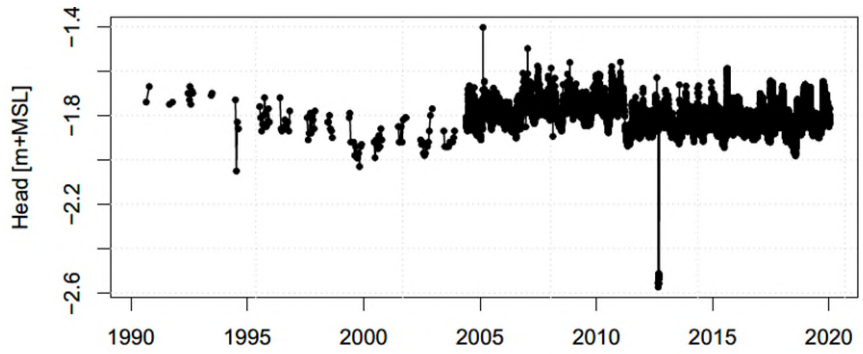

Figure S1: Example of a series with a clearly undesired outlier. The isolated low points in 2013 are caused by temporary extraction to clean the well. The series also contains smaller outliers in e.g. 1994 and 2006.

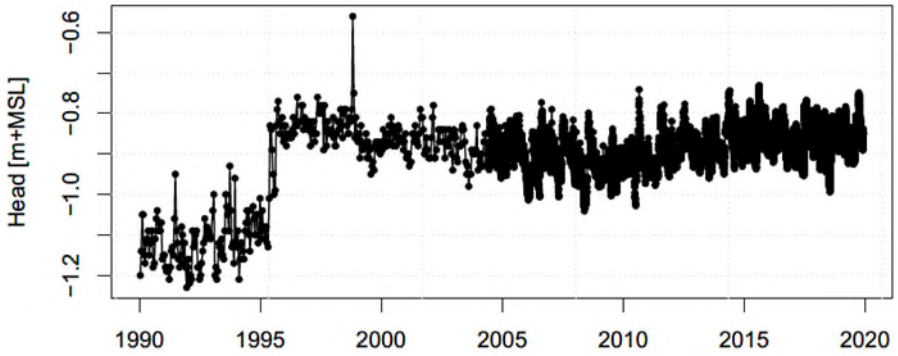

Figure S2: Example of a series with a serious long-term deviation. The level shift around 1996 is caused by changes in water management in the area. Also an outlier is visible in 1999.

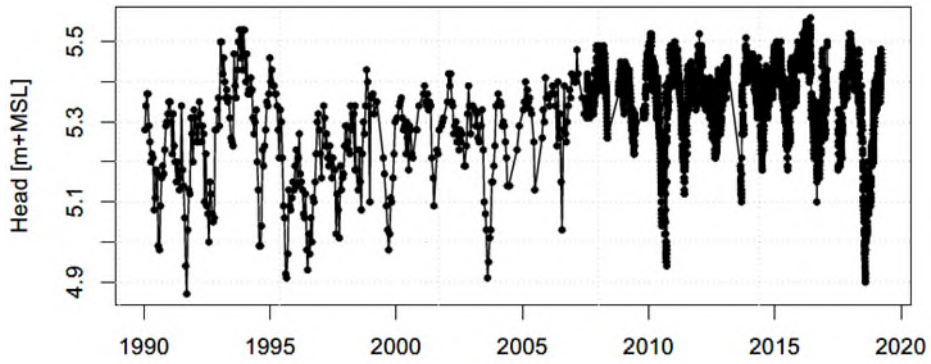

Figure S3: Example of a series with potentially unreliable long-term behaviour. There seems to be a trend from 1996 and a stabilisation after 2006, but the variations could be caused by weather variations and do not give reason to discard the series.

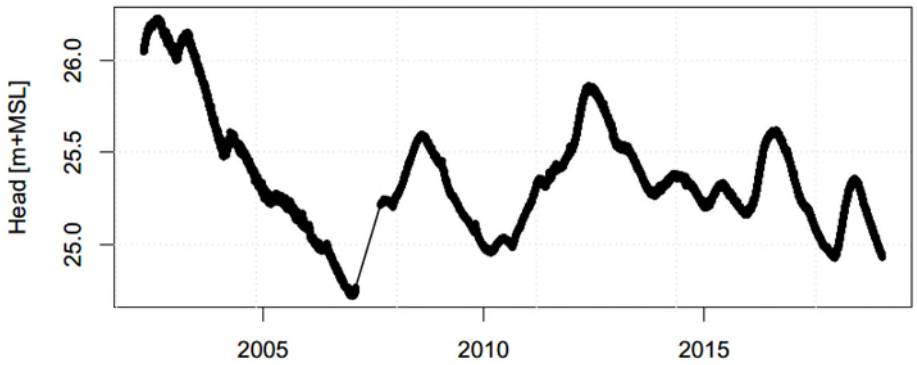

Figure S4: Example of a series with a deep groundwater table and resulting multi-year variation. Series located on the Veluwe ice-pushed ridge, mean WTD 35 m.

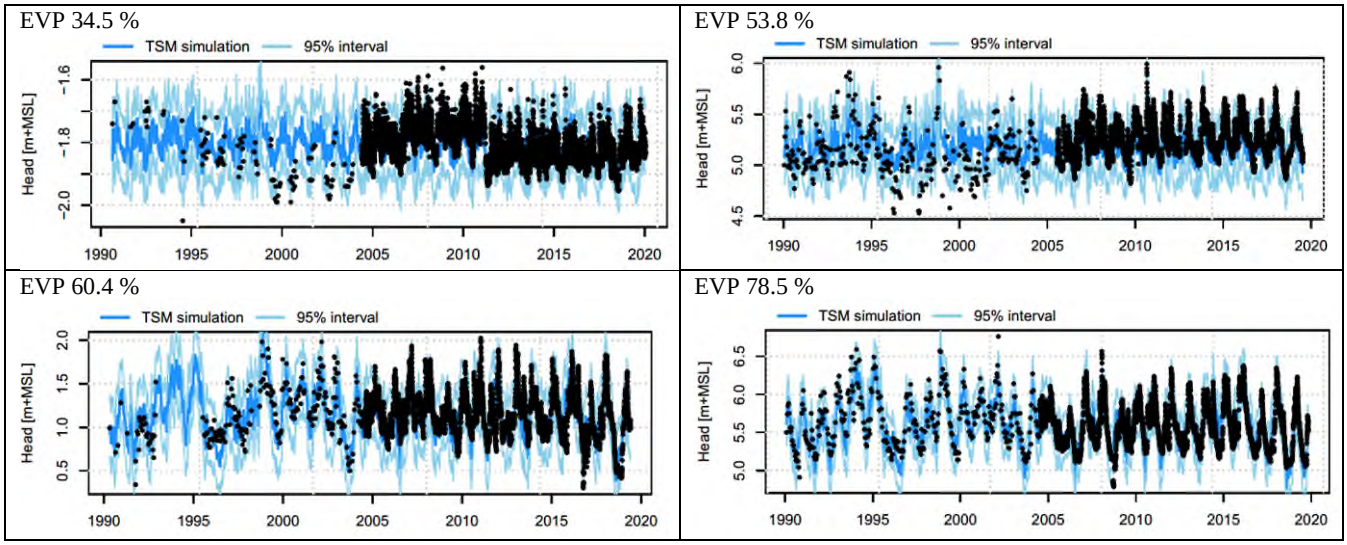

Figure S5: Examples of series with a range of EVP values (explained variance percentage).

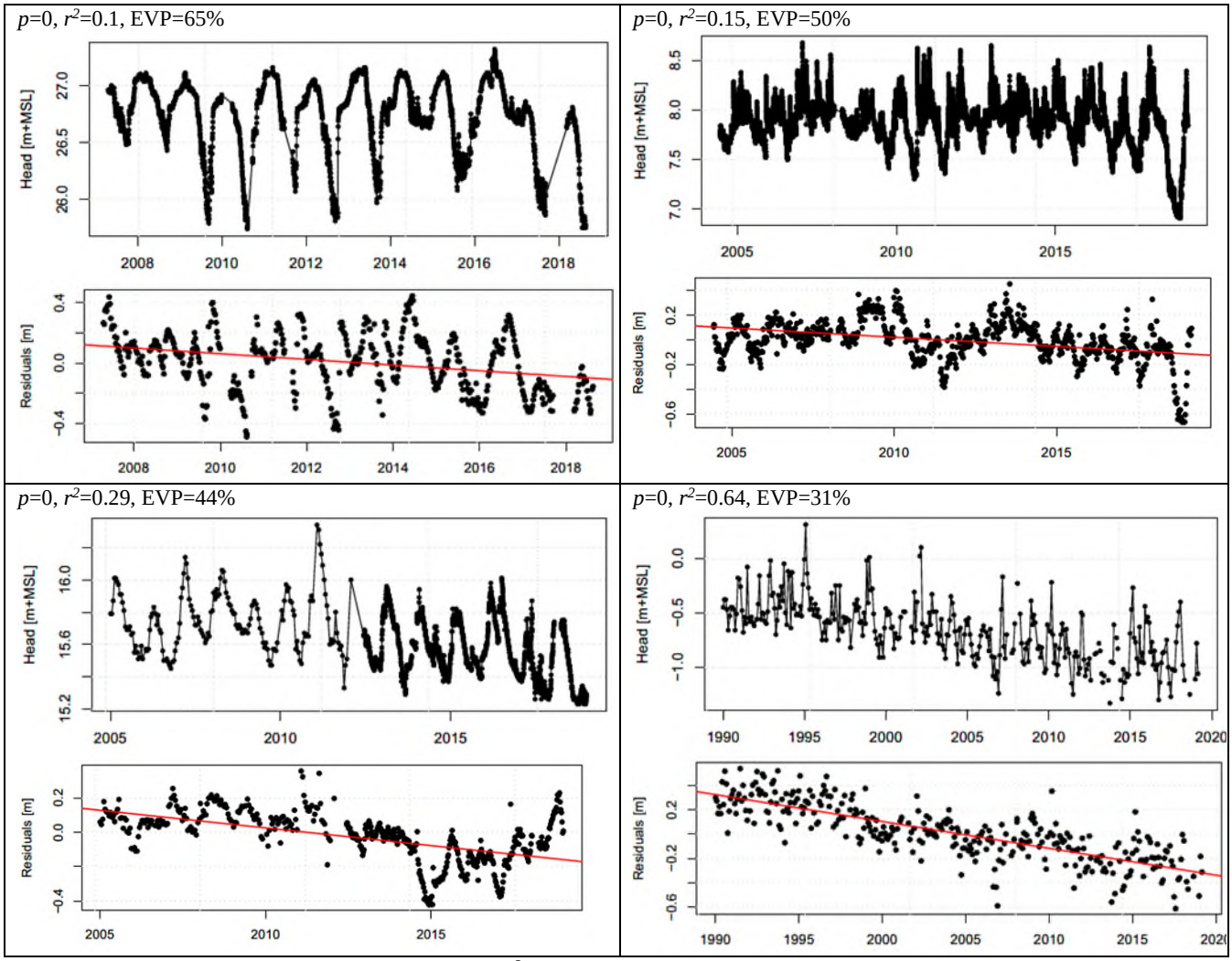

Figure S6: Examples of series with a range of trend r^2 values in the model residuals. Both the series itself and the residuals with fitted trend are shown.

S2 Validation parameter sensitivity

S2.1 Parameter sets

The validation routine was run with 21 different parameter sets (Table A2). In set 2-11, the parameters are varied individually; in set 12-19, combinations of outlier cleaning and long-term deviation identification parameters are tested; and in set 20 and 21 versions are tested with only TSM-based outlier cleaning (no basic cleaning step) and no outlier cleaning at all.

Table S1: Used parameter sets for the sensitivity test. See table 1 for explanation of the parameters.

Nr	Name	Meta Check	F_{range}	TH_{red}	n_{SD}	n_{iter}	TH_{EVP}	TH_{r2}
1	*Standard*	*Yes*	*0.2*	*0.5*	*4*	*2*	*60*	*0.15*
2	FarOutliersConservative	Yes	0.1	0.7	4	2	60	0.15
3	FarOutliersRigorous	Yes	0.2	0.4	4	2	60	0.15
4	OutliersConservative	Yes	0.2	0.5	6	2	60	0.15
5	OutliersRigorous	Yes	0.2	0.5	2	2	60	0.15
6	IterationsConservative	Yes	0.2	0.5	4	1	60	0.15
7	IterationsRigorous	Yes	0.2	0.5	4	5	60	0.15
8	EVPConservative	Yes	0.2	0.5	4	2	40	0.15
9	EVPRigorous	Yes	0.2	0.5	4	2	80	0.15
10	TrendConservative	Yes	0.2	0.5	4	2	60	0.4
11	TrendRigorous	Yes	0.2	0.5	4	2	60	0.05
12	OutliersConservative_EVPConservative	Yes	0.1	0.7	6	1	40	0.15
13	OutliersConservative_EVPRigorous	Yes	0.1	0.7	6	1	80	0.15
14	OutliersConservative_TrendConservative	Yes	0.1	0.7	6	1	60	0.4
15	OutliersConservative_TrendRigorous	Yes	0.1	0.7	6	1	60	0.05
16	OutliersRigorous_EVPConservative	Yes	0.2	0.4	2	5	40	0.15
17	OutliersRigorous_EVPRigorous	Yes	0.2	0.4	2	5	80	0.15
18	OutliersRigorous_TrendConservative	Yes	0.2	0.4	2	5	60	0.4
19	OutliersRigorous_TrendRigorous	Yes	0.2	0.4	2	5	60	0.05
20	Standard_TSMcleaningOnly	No	0	1	4	2	60	0.15
21	Standard_NoOutlierCleaning	No	0	1	100	0	60	0.15

S2.2 Test results

Outliers

Table A3 gives the full outlier cleaning performance results for all parameter sets. Cleaning of outliers in general was able to increase the EVP of series TSM models from 60 % to 62 % on average (set 1 vs 21). Also, applying a cleaning step allows for more series to be retained. The basic, range-based outlier step adds relatively little to the cleaning quality compared to the time series model-based cleaning (set 20 vs 1). The range-based cleaning step appeared useful in a small number of cases where it improved the quality of the TSM cleaning; in addition, it is a computationally cheap step and was therefore retained in the validation.

As expected from the minor effect of the far outlier cleaning, the thresholds of this step have little effect on the validation performance (set 2 and 3). Changing the TSM-based outlier cleaning thresholds does result in large effects: lowering the SD

threshold to 2·SD (set 5) causes the mean EVP (logically) to increase substantially, but also causes removal of many data points that would visually not be identified as errors (many false positives). Applying a more conservative TSM outlier cleaning with 6·SD (set 4) has the reverse effect, with many outliers not identified and more series being discarded. The number of iterations applied in the TSM outlier cleaning only slightly affected the resulting EVP of the series models. Apparently the first cycle already cleans the main outliers in most cases.

Table S2: Validation performance with regard to outliers. Left columns: number of series (of *n*=180) classified in each category. TP=outliers identified in manual and automatic validation; TN=outliers not identified in either manual or automatic validation; FP=outliers identified in automatic validation but not in manual validation; FN=outliers identified in manual but not in automated validation. Last column: percentage of series with outliers correctly cleaned.

Set	Name	Missing Data	Outliers TP	Outliers TN	Outliers FP	Outliers FN	Mean EVP	Outliers Good [%]
1	Standard	19	71	56	26	8	61.9	**79**
2	FarOutliersConservative	19	72	56	26	7	62	**80**
3	FarOutliersRigorous	19	71	56	26	8	61.9	**79**
4	OutliersConservative	19	36	92	5	28	60.6	**80**
5	OutliersRigorous	19	96	0	64	1	68.8	**60**
6	IterationsConservative	19	71	56	26	8	61.7	**79**
7	IterationsRigorous	19	71	56	26	8	62.2	**79**
8	EVPConservative	19	71	56	26	8	61.9	**79**
9	EVPRigorous	19	71	56	26	8	61.9	**79**
10	TrendConservative	19	71	56	26	8	61.9	**79**
11	TrendRigorous	19	71	56	26	8	61.9	**79**
12	OutliersConservative_ EVPConservative	19	37	92	5	27	60.5	**80**
13	OutliersConservative_ EVPRigorous	19	37	92	5	27	60.5	**80**
14	OutliersConservative_ TrendConservative	19	37	92	5	27	60.5	**80**
15	OutliersConservative_ TrendRigorous	19	37	92	5	27	60.5	**80**
16	OutliersRigorous_ EVPConservative	22	93	0	64	1	71.1	**59**
17	OutliersRigorous_ EVPRigorous	22	93	0	64	1	71.1	**59**
18	OutliersRigorous_ TrendConservative	22	93	0	64	1	71.1	**59**
19	OutliersRigorous_ TrendRigorous	22	93	0	64	1	71.1	**59**
20	Standard_ TSMcleaningOnly	19	73	54	26	8	62.1	**79**
21	Standard_ NoOutlierCleaning	19	0	110	0	51	59.9	**68**

Serious long-term deviations: discarding of series

Table A4 shows the validation performance with regard to serious long-term deviations in the series. Changing the EVP threshold for discarding series logically has a strong effect on the number of discarded series (set 8 and 9). Taking a conservative low EVP threshold appears to give a good performance on the strong deviation identification (set 8 and 16), but the number of false negatives, leading to potentially erroneous outcomes, is high. The outlier cleaning also affects the identification of long-term deviations. More rigorous outlier cleaning, especially by reducing the SD threshold, led to discarding a smaller number of series and slightly better performance on the long-term deviation identification, with conservative outlier cleaning having the reverse effect (set 2-5). However, the changes are small.

Table S3: Validation performance with regard to strong long-term deviation. Left columns: number of series (of *n*=180) classified in each category. TP=discarded in manual and automatic validation; TN=not discarded in either manual or automatic validation; FP=discarded in automatic validation but not in manual validation; FN=discarded in manual but not in automated validation. Excl deep: false positives excluding deep-GWL series. Discard number: number of series discarded of *n*=180. Last column: percentage of series with long-term strong deviation correctly identified.

Set	Name	Discard TP	Discard TN	Discard FP	Discard FP excl deep	Discard FN	Discard number	Discard Good [%]
1	Standard	32	103	26	22	0	58	**84**
2	FarOutliersConservative	32	103	26	22	0	58	**84**
3	FarOutliersRigorous	32	103	26	22	0	58	**84**
4	OutliersConservative	33	97	31	27	0	64	**81**
5	OutliersRigorous	24	115	17	13	5	41	**86**
6	IterationsConservative	32	100	29	25	0	61	**82**
7	IterationsRigorous	32	103	26	22	0	58	**84**
8	EVPConservative	17	121	12	8	11	29	**86**
9	EVPRigorous	34	42	85	73	0	119	**47**
10	TrendConservative	32	103	26	22	0	58	**84**
11	TrendRigorous	32	103	26	22	0	58	**84**
12	OutliersConservative_ EVPConservative	17	121	12	8	11	29	**86**
13	OutliersConservative_ EVPRigorous	34	39	88	74	0	122	**45**
14	OutliersConservative_ TrendConservative	33	95	33	29	0	66	**80**
15	OutliersConservative_ TrendRigorous	33	95	33	29	0	66	**80**
16	OutliersRigorous_ EVPConservative	12	129	3	1	14	15	**89**
17	OutliersRigorous_ EVPRigorous	32	76	50	41	0	82	**68**
18	OutliersRigorous_ TrendConservative	21	116	15	11	6	36	**87**
19	OutliersRigorous_ TrendRigorous	21	116	15	11	6	36	**87**
20	Standard_ TSMcleaningOnly	32	104	25	21	0	57	**84**
21	Standard_ NoOutlierCleaning	33	97	31	27	0	64	**81**

Mild long-term deviations: series to be marked as atypical

The fraction of series marked as atypical in the data is relatively small across the different parameter sets. The used r^2 threshold to identify trends affects this number. A high threshold of 0.4 results in very few atypical series being identified; a low threshold results in many false positives, as weak trends were present in a large proportion of the series. The identification of atypical series is also affected by the outliers cleaning parameters. More rigorous outlier cleaning (set 5) causes more series to be marked as atypical. Also the EVP threshold to discard series affects the number of series marked as atypical. With a high EVP

threshold discards many mildly atypical series are discarded by an insufficient EVP. A low EVP threshold brings more series into the 'atypical' category.

Table S4: Validation performance with regard to mild long-term deviations. Left columns: number of series (of *n*=180) classified in each category. TP=marked as atypical in both manual and automatic validation; TN=not marked as atypical in either manual or automatic validation; FP=marked as atypical in automatic validation but not in manual validation; FN=marked as atypical in manual but not in automated validation. Excl deep: false positives excluding deep-GWL series. Atyp number: number of series marked atypical of n=180. Last columns: percentage of series with mild long-term deviations correctly identified.

Set	Name	AtypTP	Atyp TN	Atyp FP	Atyp FP excl deep	Atyp FN	Atyp number	Atyp Good [%]
1	Standard	7	131	11	7	12	18	86
2	FarOutliersConservative	7	131	11	7	12	18	86
3	FarOutliersRigorous	7	131	11	7	12	18	86
4	OutliersConservative	7	132	10	6	12	17	86
5	OutliersRigorous	9	123	19	14	10	28	82
6	IterationsConservative	7	131	11	7	12	18	86
7	IterationsRigorous	7	131	11	7	12	18	86
8	EVPConservative	10	119	23	18	9	33	80
9	EVPRigorous	0	142	4	2	15	4	88
10	TrendConservative	2	139	5	1	15	7	88
11	TrendRigorous	8	111	31	24	11	39	74
12	OutliersConservative_ EVPConservative	10	122	20	15	9	30	82
13	OutliersConservative_ EVPRigorous	0	142	4	2	15	4	88
14	OutliersConservative_ TrendConservative	2	139	5	1	15	7	88
15	OutliersConservative_ TrendRigorous	7	113	29	22	12	36	75
16	OutliersRigorous_ EVPConservative	13	106	32	24	7	45	75
17	OutliersRigorous_ EVPRigorous	2	135	6	4	15	8	87
18	OutliersRigorous_ TrendConservative	3	131	10	5	14	13	85
19	OutliersRigorous_ TrendRigorous	10	104	35	30	9	45	72
20	Standard_ TSMcleaningOnly	8	131	11	7	11	19	86
21	Standard_ NoOutlierCleaning	8	134	8	4	11	16	88