



# Improving Linear Interpolation of Missing Hydrological Data by Applying Integrated Autoregressive Models

Tomasz Niedzielski<sup>1</sup> · Michał Halicki<sup>1</sup>

Received: 27 June 2023 / Accepted: 18 September 2023 / Published online: 3 October 2023  
© The Author(s) 2023

## Abstract

The application of linear interpolation for handling missing hydrological data is unequivocal. On one hand, such an approach offers good reconstruction in the vicinity of last observation before a no-data gap and first measurement after the gap. On the other hand, it omits irregular variability of hydrological data. Such an irregularity can be described by time series models, such as for instance the autoregressive integrated moving average (ARIMA) model. Herein, we propose a method which combines linear interpolation with autoregressive integrated model (ARI, i.e. ARIMA without a moving average part), named LinAR (available at GitHub), as a tool for inputting hydrological data. Linear interpolation is combined with the ARI model through linear scaling the ARI-based prediction issued for the no-data gap. Such an approach contributes to the current state of art in gap-filling methods since it removes artificial jumps between last stochastic prediction and first known observation after the gap, also introducing some irregular variability in the first part of the no-data gap. The LinAR method is applied and evaluated on hourly water level data collected between 2016 and 2021 (52,608 hourly steps) from 28 gauges strategically located within the Odra/Oder River basin in southwestern and western Poland. The data was sourced from Institute of Meteorology and Water Management (Poland). Evaluating the performance with over 100 million assessments in the validation experiment, the study demonstrates that the LinAR approach outperforms the purely linear method, especially for short no-data gaps (up to 12 hourly steps) and for rivers of considerable size. Based on rigorous statistical analysis of root mean square error (RMSE) – expressed (1) absolutely, (2) as percentages and (3) using RMSE error bars – the percentage improvement, understood as percentage difference between RMSE of linear and LinAR interpolations, was found to reach up to 10%.

**Keywords** Hydrology · Imputation · Water level · Linear interpolation · Autoregression

---

Tomasz Niedzielski and Michał Halicki contributed equally to this work.

---

✉ Michał Halicki  
michal.halicki2@uwr.edu.pl

Tomasz Niedzielski  
tomasz.niedzielski@uwr.edu.pl

<sup>1</sup> Department of Geoinformatics and Cartography, Faculty of Earth Sciences and Environmental Management, University of Wrocław, pl. Uniwersytecki 1, 50-137 Wrocław, Poland

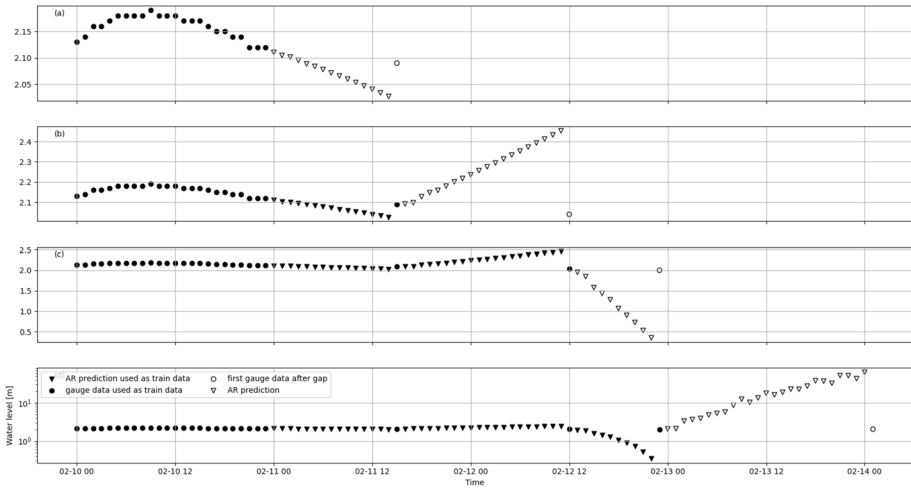
## 1 Introduction

Although the level of completeness of hydrological time series is meant to be increasing (Dixon 2010), handling missing data in hydrology still remains to be a challenge. Gaps in water-related datasets not only constrain the analysis and interpretation of historical river-flow variation but may also have a deleterious influence on hydrological models and the associated predictions (Harvey et al. 2012). Indeed, Gill et al. (2007) claimed that ignoring missing values before modelling is more detrimental for data-based models, such as those based on learning algorithms, than for physically-based approaches (Zhang and Post 2018).

In early 2010's, Harvey et al. (2012) argued that "there are currently no widely-accepted standard techniques for data infilling, either in the UK or internationally". More recently, Gao et al. (2018) provided a similar assessment of the state-of-art in the field of handling missing data in hydrology and wrote that "Imputation in hydrology has very often been done in an ad hoc manner, lacking a clear theoretical basis". Though there are attempts to develop robust gap-filling approaches (e.g. Dembélé et al. 2019; Hamzah et al. 2021; Khampuangson and Wang 2023), commonly, various statistical methods are adopted to reconstruct missing parts of a hydrograph (e.g. McCuen 2003). In hydrological sciences, however, hydrograph data reveal certain specifics – a high flow is characterized by a rising limb, a crest, and a slowly-declining recession limb (e.g. Reddy 2005). Therefore, adopting a given statistical method should not be carried out ad hoc, but it needs to account for shape of a hydrograph and/or its specific variability.

Though gap filling using hydrological models was found to have a negligible impact on the determination of trend in riverflow (Zhang and Post 2018), such an approach may provide valuable information when water level predictions are issued. Indeed, to build a lumped model based on time series methods, such as for instance vector autoregression (Niedzielski 2007; Niedzielski and Miziński 2017), continuous time series is required. As mentioned above, hydrological data reveal the specific shape, and therefore flexible non-linear approaches, such as autoregressive integrated moving average (ARIMA) or autoregressive conditional heteroscedastic (ARCH) ones (Gao et al. 2018), are needed. It has been found that the application of ARIMA models in the process of filling gaps in streamflow time series is justified, with limitation on the number of missing data (Lopes Martins et al. 2023). Furthermore, the ARIMA approach is not computationally-demanding (Ren et al. 2022), which makes it applicable in hydrological applications that require rapid data processing.

The ARIMA models, however, when used to simulate or predict variation of hydrological data within a no-data gap, may occasionally reveal instability driven by integrating a spike-rich reconstructed part produced by combining predictions with forthcoming data (Fig. 1). Instability of ARIMA models and the deterioration of their performance has been reported by Li et al. (2023) who claim that there is scarcity of research in this topic. Also, the instability in question has been reported in the hydrological context by Gui et al. (2021) who argue that the ARIMA approach belongs to a set of a few mathematical models which reveal unsatisfactory skills in simulating streamflow. Herein, we propose a new method and its Python implementation to address the problem of instability of ARIMA models used for data gap infilling in hydrology. The structure of the paper is the following: first we describe our gap-filling methodology, named hereinafter as LinAR (Section 2), second we present datasets used for validation (Section 3), third we show the results on the accuracy of our interpolation method (Section 4), fourth we discuss our findings with similar studies (Section 5), and finally we conclude the article (Section 6).



**Fig. 1** Example of unrealistic instability of water level reconstructions (triangles) based on the ARIMA model fitted to data (black circles) or data+predictions (black circles + black triangles): **a** predictions imputed in the 15-step no-data gap (triangles with white filling) along with single real data occurring after the gap (circle with white filling), **b** data+predictions+data (black symbols) from (a) are modelled and used to produce forecasts for filling the 20-step no-data gap (triangles with white filling) along with next single real data occurring after the gap (circle with white filling), **c** data+predictions+data+predictions+data (black symbols) from (b) are modelled and utilized to produce forecasts for filling the 10-step no-data gap (triangles with white filling) along with next single real data occurring after the gap (circle with white filling), **d** data+predictions+data+predictions+data+prediction+data (black symbols) from (c) are modelled and used to produce forecasts for filling the 25-step no-data gap (triangles with white filling) along with next single real data occurring after the gap (circle with white filling)

## 2 Rationale and Methods

The primary objective of this study is to introduce the LinAR approach, which combines the ARI-based imputation with linear interpolation. By doing so, the proposed method aims to address the issue of unrealistic jumps in the hydrological time series resulting from the use of the ARIMA model and imitate short-term hydrological variability in the available data.

As explained in Fig. 1 and its caption, the use of the ARIMA model for interpolating missing hydrological data may lead to unrealistic jumps between last reconstructed values in a no-data gap and first real record. Such jumps do not occur if simple linear interpolation is employed, as in the papers by Niedzielski and Miziński (2017) and Kulanuwat et al. (2021). However, such a simplistic approach introduces artificial trend-like intervals, particularly when gaps are long, potentially causing constraints in data-based lumped modelling (i.e. fitting time series models to partially linear data is not recommended). In addition, Musial et al. (2011) argue that linear interpolation has a limited potential in offering reliable estimates as it reveals tendency to underestimate real observations, since the imputation is limited by bounds associated with true data before/after the no-data gap. Lepot et al. (2017) add that overestimation is also possible while applying linear interpolation.

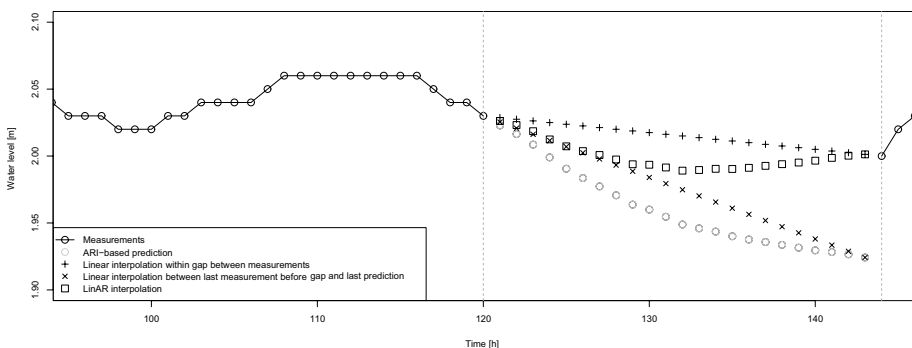
On the contrary, Gnauck (2004) claims that the linear approach performs better than the nonlinear interpolation methods in filling gaps in water quality time series. The latter finding justifies our approach, in which simple linear interpolation is used

to reconstruct a general tendency and the integrated autoregressive (ARI, i.e. ARIMA without a moving average part) model (also linear in its stochastic structure) is used to reproduce some portion of irregular variability between bounds, as postulated by Musial et al. (2011). Hence, herein we propose the combination of the ARI-based imputation with linear interpolation, both avoiding the aforementioned jumps and imitating (in short term) hydrological variability of the available data. The Python implementation of LinAR is freely available in our GitHub repository (<https://github.com/MichalHalicki4/LinAR-interpolation>).

The LinAR approach begins with the computation of autoregressive-based predictions with a lead time equal to the number of missing data points within a given no-data gap. Next, we employ linear interpolation between the last available measurement before the gap and the first measurement after the gap. Additionally, linear interpolation is applied between the last available measurement before the gap and the last autoregressive prediction. The scaling factor is then determined using the differences between the linearly interpolated values and the corresponding ARI-based predictions within the entire no-data gap (Fig. 2).

The ARI model corresponds to autoregressive (AR) model, after  $m$ -time differencing input time series  $x_t$  in order to produce residuals  $y_t$ . Firstly, each  $x_t$  is differenced to produce residuals  $\nabla x_t = x_t - x_{t-1}$ . To check stationarity, we propose to use two tests jointly, i.e. the augmented Dickey-Fuller (ADF) test for stationarity is applied to dataset and the F-test for the equality of two variances is applied to two equal-size parts of dataset (first and second half of dataset when sample size is even, and first and second half of dataset without middle value when sample size is odd). Such a two-test approach passes data with relatively constant variance, even if the ADF test fails to detect non-stationarity. If  $\nabla x_t$  is stationary, no further differencing is needed, and therefore  $\nabla x_t$  is assumed to be residuals  $y_t$  prepared for further processing. If  $\nabla x_t$  is non-stationary,  $\nabla x_t$  is again differenced to produce  $\nabla^2 x_t = \nabla(\nabla x_t) = \nabla(x_t - x_{t-1}) = x_t - 2x_{t-1} + x_{t-2}$ . If  $\nabla^2 x_t$  passes the two-test stationarity verification it becomes the residual time series  $y_t$ . In our work, we assume the significance level of 0.05.

The zero-mean autoregressive AR stochastic process  $Y_t$ , where  $y_t$  represents its trajectory, is mathematically described by the equation:



**Fig. 2** Graphical explanation of the LinAR gap imputation approach that combines ARI-prediction (gray circles) and linear interpolation (crosses)

$$Y_t = \sum_{i=1}^p a_i Y_{t-i} + Z_t, \tag{1}$$

where  $a_i$  denotes the autoregressive coefficients,  $p$  represents the order of autoregression, and  $Z_t$  signifies white noise. To produce a forecast of  $x_t$ , the prognosis of  $y_t$  is attached to residuals, and such an extended residual signal is integrated  $m$  times.

Second, linear interpolation between the last available measurement before  $n$ -step no-data gap and the first measurement after this gap is applied. Each  $k$ -th imputed value within the gap is calculated using the following expression:

$$\tilde{x}(k) = k \frac{(x_{aftergap} - x_{beforegap})}{n + 1} + x_{beforegap}, \tag{2}$$

where  $x_{aftergap}$ ,  $x_{beforegap}$  are bounds, and  $1 \leq k \leq n$ .

Third, linear interpolation between the last available measurement before  $n$ -step no-data gap and the last ARI-based prediction into step  $n$  is utilized. Each  $k$ -th value within the gap is computed along the lines of Eq. 2 as:

$$\tilde{p}(k) = k \frac{(p_{lastingap} - x_{beforegap})}{n} + x_{beforegap}, \tag{3}$$

where  $p_{lastingap}$  is the last forecasted (into  $n$ -th step) value, and  $1 \leq k \leq n - 1$ .

Fourth, both linear expressions  $\tilde{x}$  and  $\tilde{p}$  are used to produce a scaling factor, the objective of which is to correct the imputation based on ARI model. Namely, the difference:

$$\alpha(k) = \tilde{x}(k) - \tilde{p}(k) \tag{4}$$

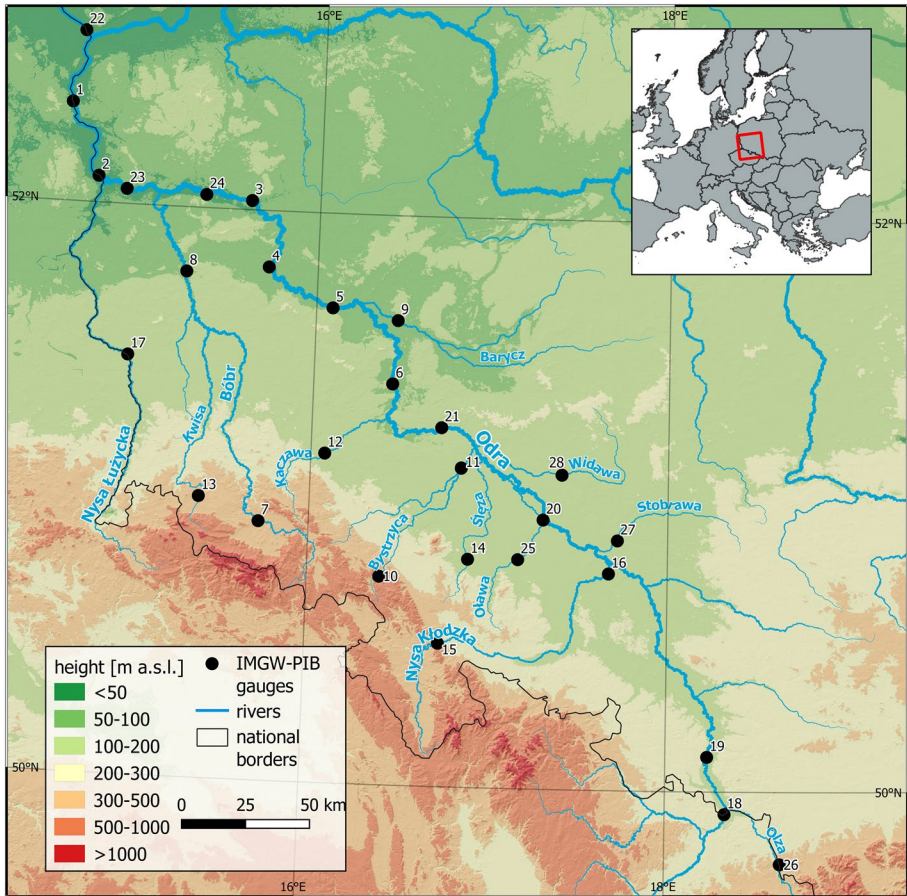
is added to every ARI-based prediction for  $k$  steps ahead, selected from a prognosis with maximum lead time of  $n$  steps and issued for the entire no-data gap (gray circles in Fig. 2 are moved up and become black squares).

### 3 Data

The imputation approach described in this paper is tested on 28 hourly water level time series, collected within the Odra/Oder River basin (SW Poland) at hydrological gauges of the Institute for Meteorology and Water Management – State Research Institute (IMGW-PIB). Spatial distribution of these gauges is presented in Fig. 3. The data span the time interval from 01/01/2016 to 31/12/2021.

The Odra/Oder River has its headwaters in Czechia. It flows northward to Poland, where it drains a considerable area of Polish territory, being the second largest river in the country. The Odra/Oder River in its middle reach becomes a transboundary river between Poland and Germany. Gauges maintained by IMGW-PIB are vertically referenced to Kronsztadt’86 vertical datum. Hence, water levels processed in this paper are values in centimetres above local zeros, the absolute heights of which are expressed in Kronsztadt’86.

Every water level time series studied in this paper contains no-data gaps (from 2 in Bardo to 318 in Ścinawa). The gaps reveal different lengths, ranging from 1 in many



**Fig. 3** Location of hydrological gauges selected for the study

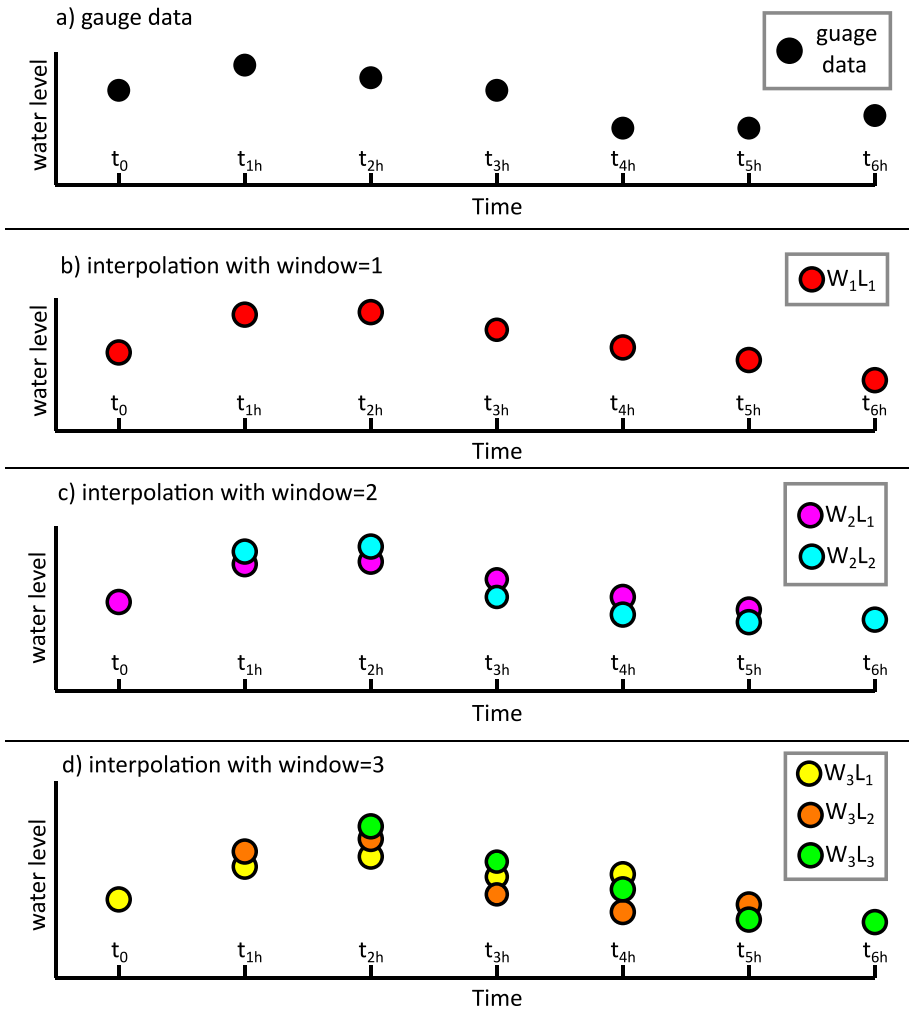
gauges to 2928 in Nowa Sól. Table 1 juxtaposes main characteristics of missing data in the above-mentioned time series.

Since reference data are needed for validation, the above-mentioned true gaps are not used to evaluate the performance of the LinAR method. To assess our approach, artificial gaps are produced so that the removed measurements are used as ground truth to compare with the interpolated values. Namely, at each hourly step selected from the studied time interval 2016–2021, we (1) artificially delete one measurement, compute interpolation and carry out the comparison, (2) artificially remove two measurements, compute interpolation of two missing data and conduct the comparison, (3) we continue the process above-described in (1) and (2) until our artificially-produced no-data gap has length of 72 steps. Literally, for each gauge we consider approx. 3,744,000 interpolations based on LinAR (52,000 hourly steps  $\times$  72). The same procedure is repeated for purely linear interpolation. To have long enough time series to fit an ARI model (we assume 5 days) and to ensure enough data for validation (72 h), we

**Table 1** Juxtaposition of gauges located along the Odra/Oder River from which data are processed in this paper. Characteristics of no-data gaps are also provided

| Gauge ID | Gauge name             | River     | River width [m] | Number of gaps | Missing data [%] | Length of gaps [h] |      |       |
|----------|------------------------|-----------|-----------------|----------------|------------------|--------------------|------|-------|
|          |                        |           |                 |                |                  | min                | max  | mean  |
| 1        | Ślubice                | Odra      | 185             | 22             | 0.5              | 1                  | 69   | 10.9  |
| 2        | Biała Góra             | Odra      | 200             | 56             | 1.8              | 1                  | 111  | 17    |
| 3        | Cigacice               | Odra      | 180             | 8              | 0.2              | 1                  | 27   | 11.1  |
| 4        | Nowa Sól               | Odra      | 100             | 13             | 8.7              | 1                  | 2928 | 353.8 |
| 5        | Głogów                 | Odra      | 85              | 13             | 0.6              | 1                  | 81   | 26.1  |
| 6        | Ścinawa                | Odra      | 70              | 319            | 4                | 1                  | 23   | 6.5   |
| 7        | Jelenia Góra           | Bóbr      | 22              | 35             | 1.3              | 1                  | 84   | 19.6  |
| 8        | Dobroszów Wielki       | Bóbr      | 35              | 12             | 0.3              | 1                  | 89   | 15    |
| 9        | Osetno                 | Barycz    | 18              | 26             | 1                | 1                  | 152  | 20.2  |
| 10       | Jugowice               | Bystrzyca | 9               | 3              | 0                | 1                  | 3    | 2.3   |
| 11       | Jarnołtów              | Bystrzyca | 12              | 6              | 0.1              | 1                  | 29   | 13    |
| 12       | Dunino                 | Kaczawa   | 9               | 13             | 0.3              | 1                  | 26   | 12.8  |
| 13       | Mirsk                  | Kwisa     | 13              | 18             | 0.5              | 1                  | 40   | 15.8  |
| 14       | Białobrzezie           | Ślęza     | 3               | 11             | 1.5              | 1                  | 636  | 70.7  |
| 15       | Bardo Śląskie          | Nysa      | 31              | 2              | 0                | 2                  | 11   | 6.5   |
| 16       | Skorogoszcz            | Nysa      | 38              | 9              | 11.3             | 1                  | 5880 | 657.7 |
| 17       | Przewóz                | Nysa      | 23              | 98             | 1.8              | 1                  | 26   | 9.5   |
| 18       | Chałupki               | Odra      | 49              | 60             | 1.4              | 1                  | 23   | 11.9  |
| 19       | Racibórz               | Odra      | 36              | 8              | 0.3              | 1                  | 39   | 18.4  |
| 20       | Miedonia               |           |                 |                |                  |                    |      |       |
| 20       | Oława                  | Odra      | 85              | 51             | 0.9              | 1                  | 28   | 8.8   |
| 21       | Brzeg Dolny            | Odra      | 70              | 8              | 0.2              | 1                  | 48   | 15.9  |
| 22       | Kostrzyn nad Odrą      | Odra      | 220             | 3              | 0.1              | 1                  | 65   | 22.3  |
| 23       | Połęcko                | Odra      | 155             | 291            | 11.6             | 1                  | 23   | 20.9  |
| 24       | Nietków                | Odra      | 140             | 5              | 0.1              | 1                  | 53   | 12    |
| 25       | Zborowice              | Oława     | 7               | 5              | 0.1              | 1                  | 19   | 9.2   |
| 26       | Cieszyn                | Olza      | 33              | 23             | 0.9              | 1                  | 23   | 21    |
| 27       | Karłowice (Wapienniki) | Stobrawa  | 6               | 10             | 0.2              | 1                  | 26   | 9.6   |
| 28       | Zbytowa                | Widawa    | 6               | 12             | 0.3              | 1                  | 26   | 12.4  |

produce buffers (from 72 h before the true no-data gap to 5 days after the gap) within which we do not interpolate. Figure 4 presents how gauge data and their interpolated equivalents correspond and how they are juxtaposed in each iteration to calculate root mean square error (RMSE).



**Fig. 4** Example sketch explaining the validation approach: **a** true water level data measured at a gauge; **b** interpolated water level when artificial no-data gaps are of length 1 (window width  $W_1$ ) and interpolation step is of length 1 (lead time  $L_1$ ), in this example  $RMSE(1, 1) = \frac{1}{7} \sum_{i=1}^7 (red_{t_{i-1}} - black_{t_{i-1}})^2$ ; **c** interpolated water level when artificial no-data gaps are of length 2 (window width  $W_2$ ) and interpolation steps are of length 1 and 2 (lead time  $L_1$  and  $L_2$ ), in this example  $RMSE(2, 1) = \frac{1}{6} \sum_{i=1}^6 (pink_{t_{i-1}} - black_{t_{i-1}})^2$  and  $RMSE(2, 2) = \frac{1}{6} \sum_{i=1}^6 (blue_{t_i} - black_{t_i})^2$ ; **d** interpolated water level when artificial no-data gaps are of length 3 (window width  $W_3$ ) and interpolation steps are of length 1, 2 and 3 (lead time  $L_1$ ,  $L_2$  and  $L_3$ ), in this example  $RMSE(3, 1) = \frac{1}{5} \sum_{i=1}^5 (yellow_{t_{i-1}} - black_{t_{i-1}})^2$ ,  $RMSE(3, 2) = \frac{1}{5} \sum_{i=1}^5 (orange_{t_i} - black_{t_i})^2$  and  $RMSE(3, 3) = \frac{1}{5} \sum_{i=1}^5 (green_{t_{i+1}} - black_{t_{i+1}})^2$



**Table 2** Accuracy of LinAR and purely linear (LIN) interpolation

| Gauge | RMSE LinAR       | RMSE LIN         | RMSE DIFF <sup>a</sup> | Improved values [%] <sup>b</sup> | Improved windows <sup>c</sup> | Average improv. [%] <sup>d</sup> |
|-------|------------------|------------------|------------------------|----------------------------------|-------------------------------|----------------------------------|
|       | min   max<br>[m] | min   max<br>[m] | min   max<br>[m]       |                                  |                               |                                  |
| 22    | 0.0047   0.0481  | 0.0047   0.0419  | -0.0065   0.0003       | 6.62                             | 3                             | 0.7                              |
| 2     | 0.0045   0.1018  | 0.0045   0.0873  | -0.0155   0.0004       | 6.01                             | 16                            | 1.4                              |
| 1     | 0.0047   0.0818  | 0.0047   0.0711  | -0.0115   0.0001       | 3.46                             | 1                             | 0.0                              |
| 3     | 0.0038   0.1325  | 0.0038   0.113   | -0.0203   0.0004       | 4.26                             | 12                            | 2.2                              |
| 23    | 0.0047   6.8727  | 0.0047   0.1097  | -6.8127   0.0018       | 10.35                            | 13                            | 3.4                              |
| 24    | 0.0046   0.1308  | 0.0046   0.1116  | -0.0203   0.0003       | 1.71                             | 3                             | 0.8                              |
| 4     | 0.004   0.1706   | 0.004   0.1435   | -0.0292   0.0006       | 4.30                             | 13                            | 3.1                              |
| 5     | 0.005   0.1958   | 0.005   0.1684   | -0.029   0.0009        | 7.80                             | 19                            | 2.9                              |
| 20    | 0.0161   0.2344  | 0.0161   0.2055  | -0.0303   0.0005       | 1.83                             | 2                             | 0.5                              |
| 6     | 0.0062   0.291   | 0.0062   0.2529  | -0.0524   0.016        | 28.81                            | 19                            | 8.7                              |
| 21    | 0.0337   8.8563  | 0.0337   0.3606  | -8.7206   0.0125       | 12.98                            | 9                             | 3.5                              |
| 18    | 0.0071   0.1864  | 0.0071   0.1674  | -0.0197   0.0002       | 2.47                             | 3                             | 0.5                              |
| 16    | 0.0079   0.1507  | 0.0079   0.1352  | -0.0163   0.0027       | 12.63                            | 5                             | 3.1                              |
| 19    | 0.0047   0.6815  | 0.0047   0.2775  | -0.5627   0.0          | 0.53                             | 3                             | 0.2                              |
| 8     | 0.015   0.4071   | 0.015   0.0975   | -0.3436   0.0019       | 19.33                            | 4                             | 1.9                              |
| 26    | 0.0111   0.7104  | 0.0111   0.1235  | -0.6611   0.0003       | 1.14                             | 2                             | 0.5                              |
| 15    | 0.0078   0.1908  | 0.0078   0.0925  | -0.0993   0.0          | 0.15                             | 1                             | 0.0                              |
| 17    | 0.0187   0.138   | 0.0187   0.1215  | -0.0584   0.0009       | 3.54                             | 2                             | 0.3                              |
| 7     | 0.0063   0.1035  | 0.0063   0.0846  | -0.0658   0.0          | 0.11                             | 1                             | 0.0                              |
| 9     | 0.0031   0.0588  | 0.0031   0.0502  | -0.009   0.0           | 0.08                             | 1                             | 0.0                              |
| 13    | 0.0079   0.1074  | 0.0079   0.0963  | -0.0113   0.0009       | 5.56                             | 7                             | 2.2                              |
| 11    | 0.0047   1.9508  | 0.0047   0.0775  | -1.9233   0.0001       | 0.80                             | 3                             | 0.4                              |
| 10    | 0.0042   0.1522  | 0.0042   0.0518  | -0.1258   0.0          | 0.04                             | 1                             | 0.0                              |
| 12    | 0.0026   0.1906  | 0.0026   0.0739  | -0.1638   0.0          | 0.19                             | 1                             | 0.0                              |
| 25    | 0.004   0.4863   | 0.004   0.1013   | -0.4444   0.0          | 0.19                             | 2                             | 0.0                              |
| 27    | 0.0035   1.0388  | 0.0035   0.0438  | -1.0236   0.0          | 0.27                             | 1                             | 0.0                              |
| 28    | 0.0044   0.1016  | 0.0044   0.0861  | -0.0393   0.0001       | 1.86                             | 4                             | 0.4                              |
| 14    | 0.0042   0.0939  | 0.0042   0.0903  | -0.0046   0.0          | 0.15                             | 1                             | 0.0                              |

<sup>a</sup>Range of differences (LIN – LinAR interpolation RMSE) for each window (W) and lead time (L) combinations

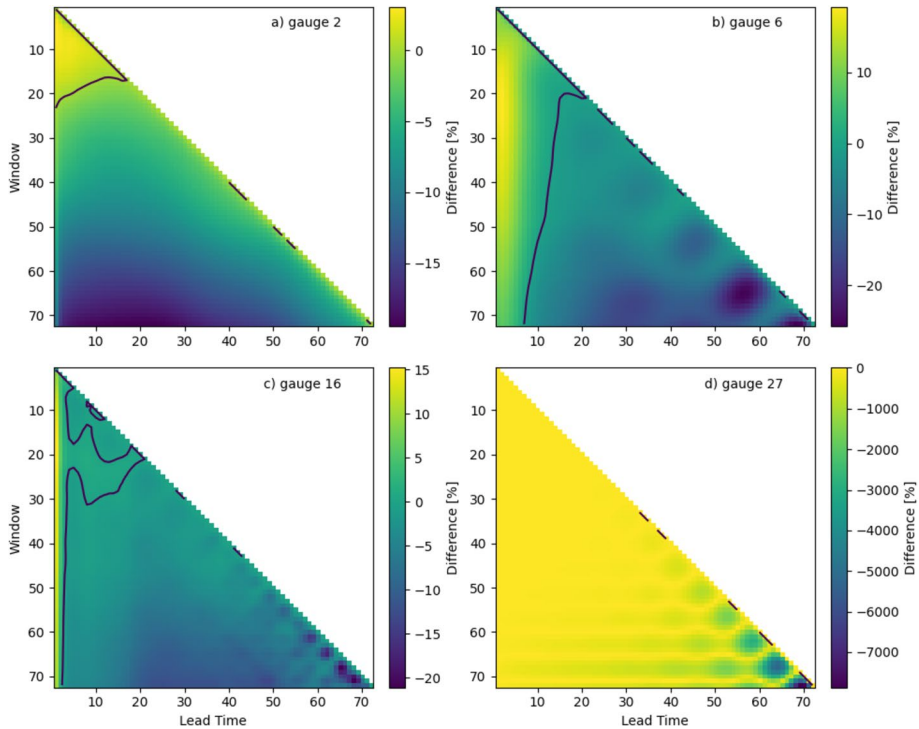
<sup>b</sup>Percentage of validation datasets (window (W) and lead time (L) combinations) for which LinAR RMSE  $\leq$  LIN RMSE

<sup>c</sup>Number of W starting from W = 1 where all LinAR RMSE  $\leq$  LIN RMSE

<sup>d</sup>Average accuracy improvement in datasets determined in previous column

## 4 Results

Although we study 28 hydrological time series, our scrutiny begins with four specific cases (Fig. 5) that allow the reader to follow complete findings juxtaposed in Table 2 and Appendix. Black contour separates interpolations for which the LinAR method performs better than the purely linear approach.



**Fig. 5** Examples of percentage differences between linear and LinAR interpolation errors (RMSE) as a function of width of no-data gap ( $W$ ) and interpolation length ( $L$ ): Odra river at Biała Góra (a), Odra river at Ścinawa (b), Nysa Kłodzka at Skorogoszcz (c), Stobrawa at Kartowice (d). Positive values indicate better performance of the LinAR approach, while negative numbers correspond to an opposite case

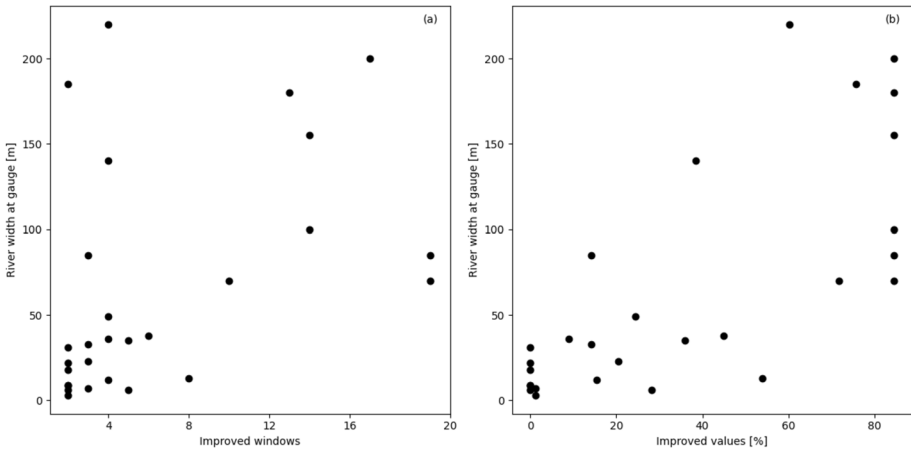
The most expected result has a graphical representation as in Fig. 5a for gauge 2 (Biała Góra, Odra, river width at gauge of 200 m, gauge within lowland). On average, in a no-data gaps of widths up to 16 ( $W = 1, 2, \dots, 16$ ), all LinAR interpolations with lengths up to 15 ( $L = 1, 2, \dots, 15$ ) are characterised by a smaller RMSE than RMSE for the linear method. However, the improvement is up to 3% of RMSE. For longer no-data gaps the linear approach is superior over the LinAR one.

A different picture is shown in Fig. 5b that corresponds to gauge 6 (Ścinawa, Odra, river width at gauge of 70 m, gauge within lowland). The improvement (LinAR in respect to purely linear) is noticed for all no-data gaps ( $W = 1, 2, \dots, 72$ ), however, only for first steps within these gaps ( $L = 1, 2, \dots, 7$ ). In this case, the improvement can exceed 15%.

An unequivocal case is presented in Fig. 5c where improvement is noticed predominantly for very small number of steps ( $L$ ), independently on widths of no-data gaps ( $W$ ). Again, for some combinations of  $L$  and  $W$  the improvement exceeds 15%.

If our tests incorrectly pass non-stationary cases, artificial periodicities occur as presented in Fig. 5d. Such highly departing values are unique, but considerable in their absolute values, therefore they meaningfully increase the overall RMSE. Fortunately, such situations are mainly reported for long no-data gaps.

It is apparent from Fig. 5 that the LinAR approach can be recommended for short no-data gaps. It is in agreement with properties of autoregressive predictions which are

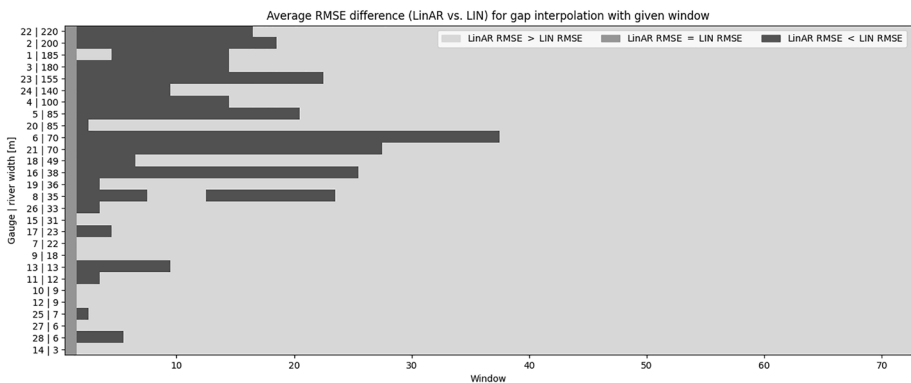


**Fig. 6** Scatter plot of interpolation improvement statistics, i.e. improved windows (a) and improved values (b), against river width. For explanation of improvement statistics see Table 2

irregular for several steps into the future and for longer lead times converge to a theoretical mean of the autoregressive process (zero for the zero-mean model).

It is apparent from Table 2 that the application of the LinAR approach leads to better results in first dozen of steps (for details, see “improved windows” column). Also, the scatter plots of “improved windows” and “improved values” against river width show that our approach is suitable for rivers of considerable size rather than for small channels (Fig. 6). It is likely that the dependence of the LinAR performance on river width is driven by less abrupt streamflow variability in wider channels. For instance, Hwang et al. (2020) claim that “[...] the river width affects flow rate significantly.” Autoregressive processes are linear and stationary time series models, and thus they have only limited potential in describing non-stationary and non-linear data (Karthikeyan and Nagesh Kumar 2013).

When averaging RMSE differences – such as those presented in Fig. 5, but for each gauge – over interpolation length (L), mean RMSE differences can be expressed as a function of



**Fig. 7** Comparison of performance of LinAR approach in respect to purely linear interpolation method. Gauges in y-axis are sorted by river width and placed in the descending order from the top of figure

width of no-data gap (Fig. 7). When gauges are ordered according to local river width, it is apparent that the LinAR approach can be recommended for sites where river is wide, and the performance is limited to a dozen of steps.

In order to estimate the recommended width of no-data gap ( $\hat{W}$ ) for which the LinAR method is superior over the purely linear approach, we make use of convex combination of improved windows in the following way:

$$\hat{W} = \sum_{g=1}^{28} W_g i_g, \quad (5)$$

where  $g$  is the index of a given gauge (Table 1),  $W_g$  is the value known in Table 2 as the improved windows, and  $i_g$  is the ratio of percentage average improvement to sum of percentage average improvement values (Table 2). When  $W_g$  and  $i_g$  values from Table 2 are imputed into Eq. 5, the estimate  $\hat{W}$  is of 12. It means that – based on the aforementioned assumptions – our approach is recommended for gaps of lengths up to 12 steps.

When we consider no-data gaps of length up to 12, we may clearly notice that percentage of improved values is meaningful (reaching even 85%), as juxtaposed in Table 3. This statistics is particularly high for gauges where river width is considerable. In addition, percentage difference in RMSE can attain as much as 10% on average (gauge 6). In contrast, there is no improvement or there exists deterioration of water level interpolation accuracy for narrow river sections (up to  $-2.5\%$ ).

Although the statistics that describe the percentage improvement of interpolation accuracy are promising, the differences in RMSE, which is a measure of water level interpolation error in metres, are very low (Tables 2 and 3). It means that the significance of improvement may be questioned. According to Kalarus et al. (2010), to say if a given mean prediction error is smaller than the other it is necessary to use error bars of mean prediction error. Therefore, to evaluate the significance of the improvement (LinAR in respect to purely linear interpolation), RMSE error bars are used following the paper by Niedzielski and Kosek (2008).

Figure 8 presents situations (combinations of  $W$  and  $L$ ) in which LinAR is significantly different from linear interpolation (for selected cases already shown in Fig. 5). One of the features of the LinAR approach is that it is identical to linear interpolation at the last interpolation point (see Fig. 2), therefore if  $W = L$  the RMSE error bar is always greater than RMSE difference (which in this case is equal to 0). It is apparent from Fig. 8 that for most  $(W,L)$  pairs the difference is statistically significant. Also, when we consider all scrutinized gauges and take an average over  $L$  (Fig. 9) the above-mentioned significance occurs more frequently than insignificance, predominantly for wide river sections under study.

## 5 Discussion

Our LinAR approach was found to provide unequivocal results, i.e. it works better for short gaps (up to 12 steps) and for rivers of considerable size, however it fails to improve purely linear interpolation in opposite cases. Such an unequivocal picture can also be found in a few papers that compare linear interpolation with various ARIMA-based interpolation approaches. In this context, our study presents a partly contradictory view on the opinion of Gnauck (2004) who argues that the linear interpolation is usually better than the

**Table 3** Accuracy of LinAR and purely linear (LIN) interpolation for windows  $\leq 12$ 

| Gauge | RMSE LinAR        | RMSE LIN          | RMSE DIFF <sup>a</sup> | Improved values [%] <sup>b</sup> | Average RMSE difference [%] <sup>c</sup> |
|-------|-------------------|-------------------|------------------------|----------------------------------|--|
|       | min   max<br>[cm] | min   max<br>[cm] | min   max<br>[cm]      |                                  |  |
| 22    | 0.47   0.78       | 0.47   0.78       | -0.0   0.03            | 60.26                            | 1.34                                     |
| 2     | 0.45   1.65       | 0.45   1.68       | 0.0   0.04             | 84.62                            | 1.82                                     |
| 1     | 0.47   1.26       | 0.47   1.27       | -0.0   0.01            | 75.64                            | 0.39                                     |
| 3     | 0.38   1.35       | 0.38   1.35       | 0.0   0.04             | 84.62                            | 2.47                                     |
| 23    | 0.47   3.31       | 0.47   3.38       | 0.0   0.16             | 84.62                            | 3.79                                     |
| 24    | 0.46   1.33       | 0.46   1.3        | -0.03   0.03           | 38.46                            | 0.48                                     |
| 4     | 0.4   2.0         | 0.4   2.01        | 0.0   0.06             | 84.62                            | 3.53                                     |
| 5     | 0.5   2.99        | 0.5   3.07        | 0.0   0.09             | 84.62                            | 3.92                                     |
| 20    | 1.61   7.43       | 1.61   7.31       | -0.12   0.05           | 14.10                            | -0.79                                    |
| 6     | 0.62   8.97       | 0.62   9.99       | 0.0   1.27             | 84.62                            | 10.20                                    |
| 21    | 3.37   23.19      | 3.37   23.16      | -0.07   1.19           | 71.79                            | 3.47                                     |
| 18    | 0.71   4.19       | 0.71   4.18       | -0.02   0.02           | 24.36                            | 0.02                                     |
| 16    | 0.79   5.28       | 0.79   5.25       | -0.06   0.24           | 44.87                            | 2.07                                     |
| 19    | 0.47   4.84       | 0.47   4.75       | -0.1   0.0             | 8.97                             | -0.85                                    |
| 8     | 1.5   5.38        | 1.5   5.32        | -0.12   0.18           | 35.90                            | 0.62                                     |
| 26    | 1.11   3.9        | 1.11   3.73       | -0.18   0.03           | 14.10                            | -2.40                                    |
| 15    | 0.78   2.77       | 0.78   2.7        | -0.08   0.0            | 0.00                             | -1.81                                    |
| 17    | 1.87   3.52       | 1.87   3.5        | -0.04   0.08           | 20.51                            | -0.09                                    |
| 7     | 0.63   2.35       | 0.63   2.3        | -0.06   0.0            | 0.00                             | -1.59                                    |
| 9     | 0.31   1.25       | 0.31   1.24       | -0.01   0.0            | 0.00                             | -0.85                                    |
| 13    | 0.79   3.24       | 0.79   3.17       | -0.07   0.09           | 53.85                            | 1.21                                     |
| 11    | 0.47   2.35       | 0.47   2.29       | -0.08   0.01           | 15.38                            | -1.47                                    |
| 10    | 0.42   2.28       | 0.42   2.24       | -0.05   0.0            | 0.00                             | -1.16                                    |
| 12    | 0.26   1.18       | 0.26   1.16       | -0.04   0.0            | 0.00                             | -2.42                                    |
| 25    | 0.4   2.2         | 0.4   2.14        | -0.07   0.0            | 1.28                             | -1.78                                    |
| 27    | 0.35   1.15       | 0.35   1.1        | -0.05   0.0            | 0.00                             | -1.76                                    |
| 28    | 0.44   2.79       | 0.44   2.75       | -0.04   0.01           | 28.21                            | -0.22                                    |
| 14    | 0.42   2.73       | 0.42   2.7        | -0.05   0.0            | 1.28                             | -1.11                                    |

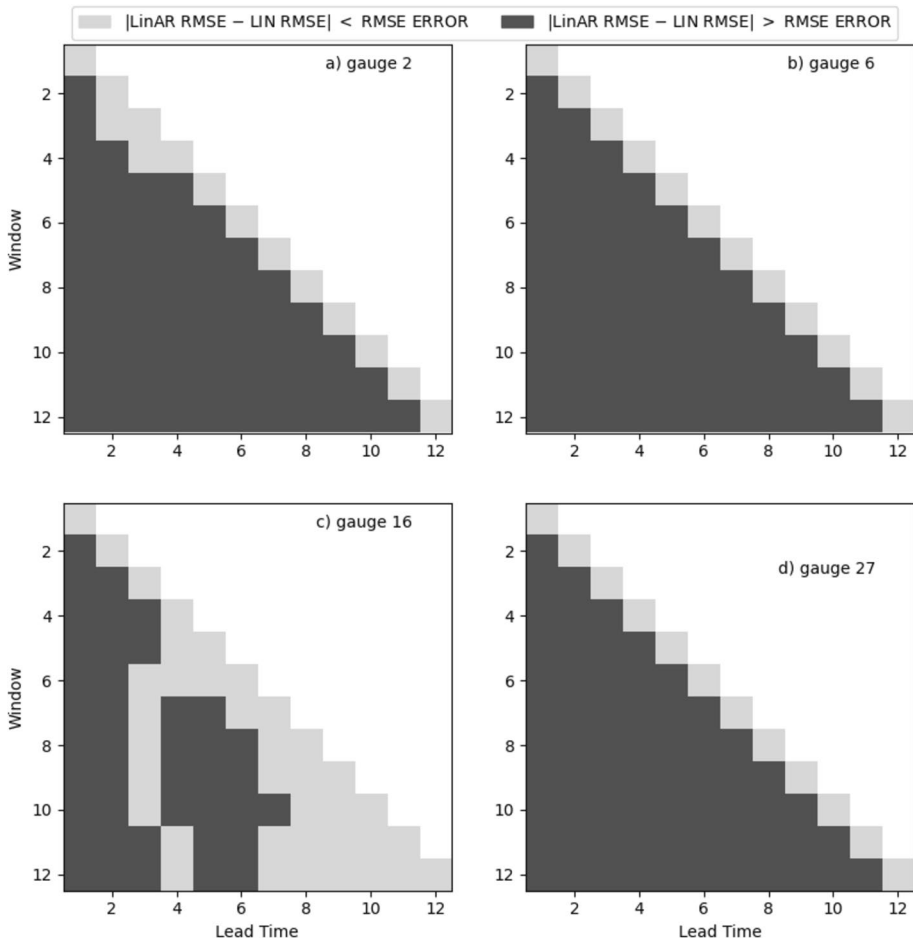
<sup>a</sup>Range of differences (LIN – LinAR interpolation RMSE) for each window (W) and lead time (L) combinations

<sup>b</sup>Percentage of validation datasets (window (W) and lead time (L) combinations) for which LinAR RMSE  $\leq$  LIN RMSE

<sup>c</sup>Percentage difference between RMSE of LIN and LinAR interpolations

non-linear one in environmental studies. Based on our experiments, similar conclusion can be drawn, however, only for long gaps and small streams. In addition, Ponkina et al. (2021) claim that the ARMA model, compared to simple linear interpolation of soil temperature and soil moisture data, was significantly better.

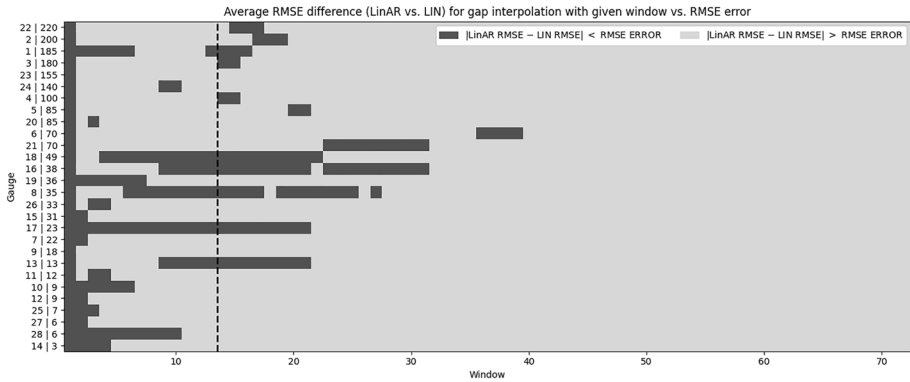
Our estimate (maximum length of gaps = 12) is in agreement with findings presented by Lopes Martins et al. (2023) who argued that the ARIMA method can be utilized for filling riverflow data gaps with lengths up to 15 steps. Such a number is explained by high autocorrelation, typical for ARIMA models, which is responsible for predictability (Tigabu et al. 2018). It should be noted that although the agreement between lengths of no-data



**Fig. 8** Significance of accuracy difference as a function of width of no-data gap ( $W$ ) and interpolation length ( $L$ ) at the selected gauges: Odra river at Biała Góra (a), Odra river at Ścinawa (b), Nysa Kłodzka at Skorogoszcz (c), Stobrawa at Kartowice (d). The significance determination was based on RMSE error bars, computed following Niedzielski and Kosek (2008)

gaps is considerable (12 vs. 15 steps), the sampling intervals were different, i.e. daily data (Lopes Martins et al. 2023) and hourly data (this study). In case of the ARIMA application, autocorrelation at a given lag (number of steps) does not directly depend on how long the step size is. For aggregated time series (e.g. daily averages), we can expect smaller hydrograph variability, better predictability and therefore good LinAR performance. Also, the superior ARIMA performance for gaps up to 12 steps has been shown in a problem of filling gaps in soil temperature and soil moisture data (Ponkina et al. 2021). Indeed, when gaps are classified into intervals (1–12, 13–24, 25–50 h), the best results correspond to the first class. More general observation has been made by Ren et al. (2022) who claim that the relative error of ARIMA interpolation increases along with length of no-data gap.

According to Ren et al. (2022), ARIMA models do not perform well when reconstructing abruptly changing water-related time series. This is in agreement with a general theory of ARIMA-based forecasting (Cholette 1982). Our scrutiny shows that the 12-step LinAR



**Fig. 9** Significance of accuracy differences as a function of  $W$ , averaged over  $L$ . Gauges in  $y$ -axis are sorted by river width and placed in the descending order from the top of figure

interpolation works on rivers of considerable size and is not recommended for small (and often mountainous) rivers. The riverflow fluctuations are usually higher and more abrupt on the latter streams, as confirmed for the study area by Sen and Niedzielski (2010), which forms a constraint in ARIMA-based modelling.

Discussion should also be developed on cases when our LinAR approach occasionally fails for long no-data gaps (see Fig. 5d). Ren et al. (2022) argue that the ARIMA approach used for no-data gap filling in hydrology produces outliers of large positive/negative errors for gap of lengths 48 and 72 hourly steps. Herein, our maximum interpolation window is of 72 steps (hours) and, as presented in Fig. 5d and explained in Section 4, for long no-data gaps the LinAR interpolation departs from reference data. Such an effect is interpreted in the light of occasional non-stationarity (if statistical tests incorrectly pass such signals to LinAR interpolation).

Recently, Khampungson and Wang (2023) presented a new method for gap infilling in hydrologic time series and compared it with five other interpolation techniques. The linear approach was found to perform best at gauges where riverflow did not reveal strong periodic patterns. Our time series are free of harmonic changes, and with LinAR we are able to perform even better.

Some national providers of hydrological data may publish raw measurements, and in such cases the LinAR approach may serve as one of potential methods for gap infilling. Although it is more complex than pure linear interpolation, it is computationally efficient and can be performed on a standard PC. For example, on a laptop with an Intel Core i7-7700HQ processor (2.80 GHz) and 16 GB RAM, a six-year hourly time series with 10% missing data (56,209 data points and 5621 gaps) is interpolated in about 13.7 s (average of 100 simulations).

## 6 Conclusions

We developed the LinAR approach (combination of linear interpolation and integrated autoregressive model) for interpolating riverflow data, checking its performance for no-data gaps ranging from 1 to 72 hourly steps. We analysed 28 water level gauges at which hourly water level data were collected between 2016 and 2021 (over 52,000 hourly steps for

each gauge). In the iterative manner, we artificially removed known measurement data to enable a reliable validation of the LinAR method. This resulted in determining RMSE values of interpolation for each length of no-data gap (1, 2, ..., 72), also considering LinAR-based predictions of various lengths (from 1 to gap size). All in all, we scrutinized over 100 million assessments ( $\approx 28 \times 52000 \times 72$ ).

Taking into account our findings and discussing them with recent papers on the ARIMA-based interpolation in hydrology, the LinAR approach may be considered useful when filling short (up to 12 hourly steps) no-data gaps in water level time series collected at rivers of considerable size. Although the superior performance of the LinAR method over purely linear interpolation has been expressed as percentage improvement (up to 10%), differences between absolute errors were very small. However, using statistical inference approaches (e.g. Niedzielski and Kosek 2008; Kalarus et al. 2010; Ren et al. 2022), they were found to be significantly more accurate than the purely linear interpolation method.

Although the LinAR method reveals a considerable potential in gap infilling in hydrological time series, it also has some limitations. First, though the LinAR implementation includes statistical tests for stationarity, they may occasionally fail, passing non-stationary signal to further analysis. This can lead to unrealistic interpolated values, especially for long no-data gaps. Thus, we recommend to use LinAR for filling short gaps. Second, if a hydrograph is highly irregular the LinAR approach may lead to unsatisfactory interpolation, because the autoregressive model reveals mediocre skills when modelling highly variable stochastic signals (usually for small rivers). Third, the absolute RMSE improvement (LinAR vs. linear interpolation) is small (but statistically significant), which may be evaluated by practitioners to have negligible effect on refining the entire process of gap imputation.

The main takeaway from the study is that the proposed LinAR approach can improve the linear interpolation of missing data in hydrology. Although LinAR uses a stochastic model, it is computationally efficient, and therefore can be utilized in practical or operational no-data gap infilling. The limitations of the approach can be mitigated by carrying further studies in: (1) better detection of non-stationarities, (2) employing various stochastic methods (other than ARI) in combination with linear interpolation.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s11269-023-03625-7>.

**Acknowledgements** The research presented in this paper has been carried out in frame of the project no. 2020/38/E/ST10/00295 within the Sonata BIS programme of the National Science Centre, Poland, as well as in frame of the Doctoral School of the University of Wrocław, Poland. Hydrological data are acquired from Institute for Meteorology and Water Management – State Research Institute (IMGW-PIB).

**Author Contributions** All authors contributed to the study conception and design. Tomasz Niedzielski conceived the study, wrote the majority of the manuscript, carried out data analysis, and produced figures. Michał Halicki developed scripts, carried out data analysis, produced figures, and contributed to the manuscript writing. All authors read and approved the final manuscript.

**Funding** This work was supported by National Science Centre, Poland within the Sonata BIS programme (Grant number 2020/38/E/ST10/00295). Michał Halicki acknowledges the funding within the Doctoral School of the University of Wrocław, Poland.

**Availability of Data and Material** The Python implementation of the LinAR method is freely available on GitHub (<https://github.com/MichalHalicki4/LinAR-interpolation>).



## Declarations

**Ethical Approval** Not applicable.

**Consent to Participate** Not applicable.

**Consent to Publish** Not applicable.

**Conflict of Interest** The authors have no relevant financial or non-financial interests to disclose.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Cholette PA (1982) Prior information and ARIMA forecasting. *J Forecast* 1:375–383. <https://doi.org/10.1002/for.3980010405>
- Dembélé M, Oriani F, Tumbulto J, Mariéthoz G, Schaeffli B (2019) Gap-filling of daily streamflow time series using direct sampling in various hydroclimatic settings. *J Hydrol* 569:573–586. <https://doi.org/10.1016/j.jhydrol.2018.11.076>
- Dixon H (2010) Managing national hydrometric data: from data to information. In: Servat E, Demuth S, Dezetter A, Daniell T (eds) *Global Change: Facing Risks and Threats to Water Resources*. Wallingford, UK, IAHS Press, pp 451–458. (IAHS Publication, 340)
- Gao Y, Merz C, Lischeid G, Schneider M (2018) A review on missing hydrological data processing. *Environ Earth Sci* 77:47. <https://doi.org/10.1007/s12665-018-7228-6>
- Gill MK, 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>
- Gnauck A (2004) Interpolation and approximation of water quality time series and process identification. *Anal Bioanal Chem* 380(3):484–492. <https://doi.org/10.1007/s00216-004-2799-3>
- Gui H, Wu Z, Zhang C (2021) Comparative study of different types of hydrological models applied to hydrological simulation. *Clean Soil Air Water* 49. <https://doi.org/10.1002/clen.202000381>
- Hamzah FB, Mohd Hamzah F, Mohd Razali SF, Samad H (2021) A comparison of multiple imputation methods for recovering missing data in hydrological studies. *Civ Eng J* 7:1608–1619. <https://doi.org/10.28991/cej-2021-03091747>
- Harvey CL, Dixon H, Hannaford J (2012) An appraisal of the performance of data-infilling methods for application to daily mean river flow records in the UK. *Hydrol Res* 43(5):618–636
- Hwang JH, Maeng SJ, Kim HS, Lee SW (2020) Analysis of river bed variation using SSARR and RMA-2 models. *Smart Water* 5:1. <https://doi.org/10.1186/s40713-019-0019-8>
- Kalarus M, Schuh H, Kosek W, Akyilmaz O, Bizouard Ch, Jovanović B, Kumakshev S, Kutterer H, Mendes Cerveira PJ, Pasynek S, Zotov L (2010) Achievements of the Earth orientation parameters prediction comparison campaign. *J Geod* 84:587–596. <https://doi.org/10.1007/s00190-010-0387-1>
- Karthikeyan L, Nagesh Kumar D (2013) Predictability of nonstationary time series using wavelet and EMD based ARMA models. *J Hydrol* 502:103–119. <https://doi.org/10.1016/j.jhydrol.2013.08.030>
- Khampungson T, Wang W (2023) Novel methods for imputing missing values in water level monitoring data. *Water Resour Manage* 37:851–878. <https://doi.org/10.1007/s11269-022-03408-6>
- Kulanuwat L, Chantrapornchai C, Maleewong M, Wongchaisuwat P, Wimala S, Sarinnapakorn K, Boonyaroonnet S (2021) Anomaly detection using a sliding window technique and data imputation with machine learning for hydrological time series. *Water* 13(13):1862. <https://doi.org/10.3390/w13131862>

- Lepot M, Aubin JB, Clemens FH (2017) Interpolation in time series: An introductory overview of existing methods, their performance criteria and uncertainty assessment. *Water* 9(10):796. <https://doi.org/10.3390/w9100796>
- Li Y, Wu K, Liu J (2023) Self-paced ARIMA for robust time series prediction. *Knowl-Based Syst* 269:110489. <https://doi.org/10.1016/j.knosys.2023.110489>
- Lopes Martins L, Martins WA, Rodrigues ICDA, Freitas Xavier AC, Moraes JFLD, Blain GC (2023) Gap-filling of daily precipitation and streamflow time series: a method comparison at random and sequential gaps. *Hydrol Sci J* 68:148–160. <https://doi.org/10.1080/02626667.2022.2145200>
- McCuen RH (2003) Modeling hydrologic change: statistical methods. CRC Press, pp 456
- Musial JP, Verstraete MM, Gobron N (2011) Comparing the effectiveness of recent algorithms to fill and smooth incomplete and noisy time series. *Atmos Chem Phys* 11(15):7905–7923. <https://doi.org/10.5194/acp-11-7905-2011>
- Niedzielski T (2007) A data-based regional scale autoregressive rainfall-runoff model: a study from the Odra River. *Stoch Environ Res Risk Assess* 21:649–664
- Niedzielski T, Kosek W (2008) Prediction of UT1-UTC, LOD and AAM  $\chi_3$  by combination of least-squares and multivariate stochastic methods. *J Geod* 82:83–92. <https://doi.org/10.1007/s00190-007-0158-9>
- Niedzielski T, Miziński B (2017) Real-time hydrograph modelling in the upper Nysa Kłodzka river basin (SW Poland): a two-model hydrologic ensemble prediction approach. *Stoch Environ Res Risk Assess* 31:1555–1576
- Ponkina E, Illiger P, Krotova O, Bondarovich A (2021) Do ARMA models provide better gap filling in time series of soil temperature and soil moisture? The case of Arable Land in the Kulunda Steppe, Russia. *Land* 10:579. <https://doi.org/10.3390/land10060579>
- Reddy PJR (2005) A text book of hydrology. Firewall Media, pp 530
- Ren H, Cromwell E, Kravitz B, Chen X (2022) Technical note: using long short-term memory models to fill data gaps in hydrological monitoring networks. *Hydrol. Earth Syst Sci* 26:1727–1743. <https://doi.org/10.5194/hess-26-1727-2022>
- Sen AK, Niedzielski T (2010) Statistical characteristics of riverflow variability in the Odra River Basin, Southwestern Poland. *Pol J Environ Stud* 19:387–397
- Tigabu TB, Hörmann G, Wagner PD, Fohrer N (2018) Statistical analysis of rainfall and streamflow time series in the Lake Tana Basin. *J Water Clim Chang, Ethiopia*. <https://doi.org/10.2166/wcc.2018.008>
- Zhang Y, Post D (2018) How good are hydrological models for gap-filling streamflow data? *Hydrol Earth Syst Sci* 22:4593–4604. <https://doi.org/10.5194/hess-22-4593-2018>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.