Supplement for “Unfolding the relationship between seasonal forecast skill and value in hydropower production: A global analysis”

Donghoon Lee[1,2], Jia Yi Ng[3], Stefano Galelli[3], and Paul Block[1]

[1]Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, Wisconsin, USA
[2]Climate Hazards Center, Department of Geography, University of California, Santa Barbara, California, USA
[3]Pillar of Engineering Systems and Design, Singapore University of Technology and Design, Singapore

Correspondence: Stefano Galelli (stefano_galelli@sutd.edu.sg)

1 Introduction

This supplement includes 5 text sections (Texts S1–S5), 3 figures (Figures S1–S3), and 5 tables (Tables S1–S5). Text S1 describes the DEM-based procedure adopted for estimating the elevation-area-storage curve for a subset of ∼200 reservoirs, while Figure S1 illustrates the impact of the bathymetry estimation procedure on the I_{PF} and I_{DF} values attained by these reservoirs. Text S2 provides the definition of mean squared error skill score (MSESS) and Gerrity skill score (GSS), while Figure S2 illustrates their values. Text S3 provides further explanation on the weekly operations of a subset of hydropower dams. Text S4–S5 and Table S1–S5 provide additional details on the regression modelling exercise used to characterize the relationship between reservoir design specifications, forecast skill, and forecast value. Figure S3 illustrates the correlation between inflow and the seven potential predictors for each of the 735 dams. All simulations results analyzed in this study are available on HydroShare at http://www.hydroshare.org/resource/ca365ffb1a1f49df8b77e393be965fd8

Text S1. The bathymetry of each reservoir is estimated using the 90-m resolution digital elevation model (DEM), flow direction, and upstream area retrieved from the Multi-Error-Removed Improved-Terrain (MERIT) – Hydro dataset (Yamazaki et al., 2019). The methodology generally follows Vu et al. (2021). First, we isolate the DEM data with the contour corresponding to maximum water level and dam crest line. Then, for each 1-m elevation change in the DEM, we calculate the corresponding water surface area. Using these data on elevation and area, we calculate the storage volume for each 1-m elevation increment (using a trapezoidal approximation), ultimately resulting in the elevation-area-storage (EAS) curve. After having estimated the EAS curves of all 735 dams, we select the 203 reservoirs with errors within 10% and 20% of the maximum dam height and maximum storage capacity reported in the GRanD database, respectively. On average, the 203 dams show a 2.5% difference in maximum capacity and a 5.5% difference in maximum dam height.

Text S2. The mean squared error skill score (MSESS) is a deterministic skill score that compares the MSE of prediction model and climatology. It is defined as follows (Wilks, 2011):

$$MSESS = \left(1 - \frac{MSE_{pred}}{MSE_{clim}}\right), \qquad (1)$$

where MSE_{pred} and MSE_{clim} are the MSE associated to the predictive model and climatological mean prediction, respectively. The perfect score of the MSESS is 1, while a value equal to 0 indicates that the model skill is equal to that of the climatology.

The Gerrity skill score (GSS) is a multi-categorical skill score that rewards correct predictions in rarer categories. The GSS is calculated as follows:

$$GSS = \sum_{i=1}^{3}\sum_{j=1}^{3} p_{ij} s_{ij}, \qquad (2)$$

where p_{ij} is the joint probability of inflow in each category (i, j) of a contingency table (3 x 3 in this study) and s_{ij} is a scoring weight to yield more or less credits based on the frequency of the category (Wilks, 2011). The three categories correspond to the upper, middle, and lower thirds of the inflow observed in the period 1958–2000. The GSS ranges from -1 to 1, where a value of 1 represents a perfect forecast and a value of 0 means no predictive skill (compared to the climatology).

The MSESS and GSS values calculated during the validation process are spatially illustrated in Figure S2. In general, models with a shorter lead-time present higher skills. As indicated in Section 3.1, the climatological mean prediction is applied instead when the MSESS or GSS value is less than 0 (e.g., yellow dots in Figure S2 panel (a)). This occurs 27% and 37% of the time (for MSESS and GSS, respectively) across all MP models.

Text S3. We perform reservoir operations at the weekly scale for 94 dams with time-to-fill (ratio of storage capacity to mean monthly inflow) lower than 2 months. To do so, we first obtain the daily inflow time series from 1958 to 1967 from the Water and Global Change (WATCH) 20$^{\text{th}}$ century Model Output (Weedon et al., 2011), generated by the global hydrological model

WaterGAP (Alcamo et al., 2003). We then aggregate the daily inflow into weeks of 7-8 days (depending on the number of days in the month) such that each month is represented by 4 weeks. We then disaggregate each monthly inflow (from 1958 to 2000) into 4 weekly inflows using the k-nearest neighbors algorithm (Nowak et al., 2010). The neighbors are identified from the 1958-1967 weekly flows. For each monthly inflow (1958 to 2000), one of the k-nearest neighbors is resampled using a weight metric that gives higher weight to the closest neighbour. The monthly inflow is then disaggregated according to the weekly proportion vector corresponding to the selected neighbor. The disaggregation of each month's inflow is independent of inflow from previous or future months. Forecast inflows are also disaggregated into weekly inflows such that the forecast horizon for forecast-informed schemes is equal to 28 weeks.

Text S4. To explain the impact of dam design specifications on the value of perfect forecasts, we first define 40 potential predictors for each dam. These predictors relate to the dam design specifications (e.g., storage capacity, maximum turbine release rate), inflow characteristics (e.g., mean and standard deviation of monthly inflow), or a combination thereof (e.g., ratio of capacity to inflow, or time-to-fill). Next, we find the correlation between these variables and I_{PF}. Table S1 lists the five variables that have correlation (absolute value) with I_{PF} greater than 0.25. Since the first three variables are correlated (all related to hydraulic head and depth), we only use x_{depth} in the modelling exercise. We then want to find design specifications that contribute to forecast value substantially. Linear regression is not best suited for this purpose, since predictors would explain small differences in I_{PF} values in which we are not interested. Hence, we perform a logistic regression exercise and divide the dams into two broad groups using the mean I_{PF} value across all dams as a cut-off. The cross-validation performance attained by logistic regression models using different combinations of predictors is shown in Table S2. We select the model with predictors x_{depth} and x_{fill} and explain it in further details in the main paper. In Table S3, we show the modelling results obtained when adopting a different threshold to divide the dams into the two groups.

Text S5. To explain the impact of forecast skills and design specifications on the value of realistic forecasts, we use the same 40 potential predictors described in Text S2 and another 23 variables characterizing the forecast skill of each inflow forecast model (Dawson et al., 2007). Similar to the previous analysis, we then calculate the correlation between these variables and the performance gain I. Table S4 shows the 14 variables that have correlation (absolute value) with I greater than 0.25. Seven of these variables are dam or inflow related, while the other seven reflect forecast skill. With these variables, we then perform an exhaustive search to find the best linear regression model with 2-to-5 predictors. The results are shown in Table S5. The model with two variables is explained in greater details in the main paper. The model with five variables includes both x_{high} and x_{hurst}, suggesting that the persistence of inflow is correlated to the value of realistic forecasts.

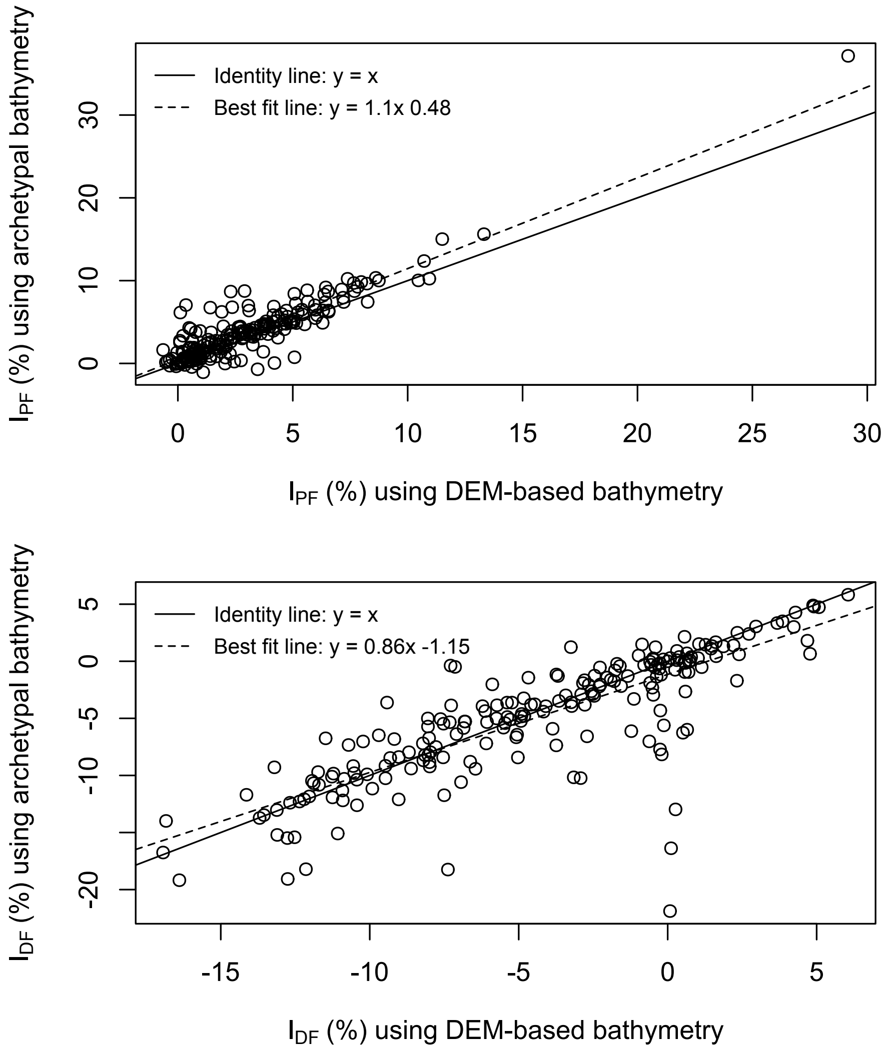

Figure S1. Comparison of performance measures (I_{PF} and I_{DF}) for 203 dams whose bathymetry (elevation-area-storage curve) is estimated using two different approaches. The first method assumes an archetypal reservoir shape (Kaveh et al., 2013), while the second method estimates the bathymetry from a high-resolution global hydrography dataset (see Text S1). This comparison is aimed to ensure that Kaveh's method provides a reasonable estimate of the bathymetry for the remaining reservoirs.

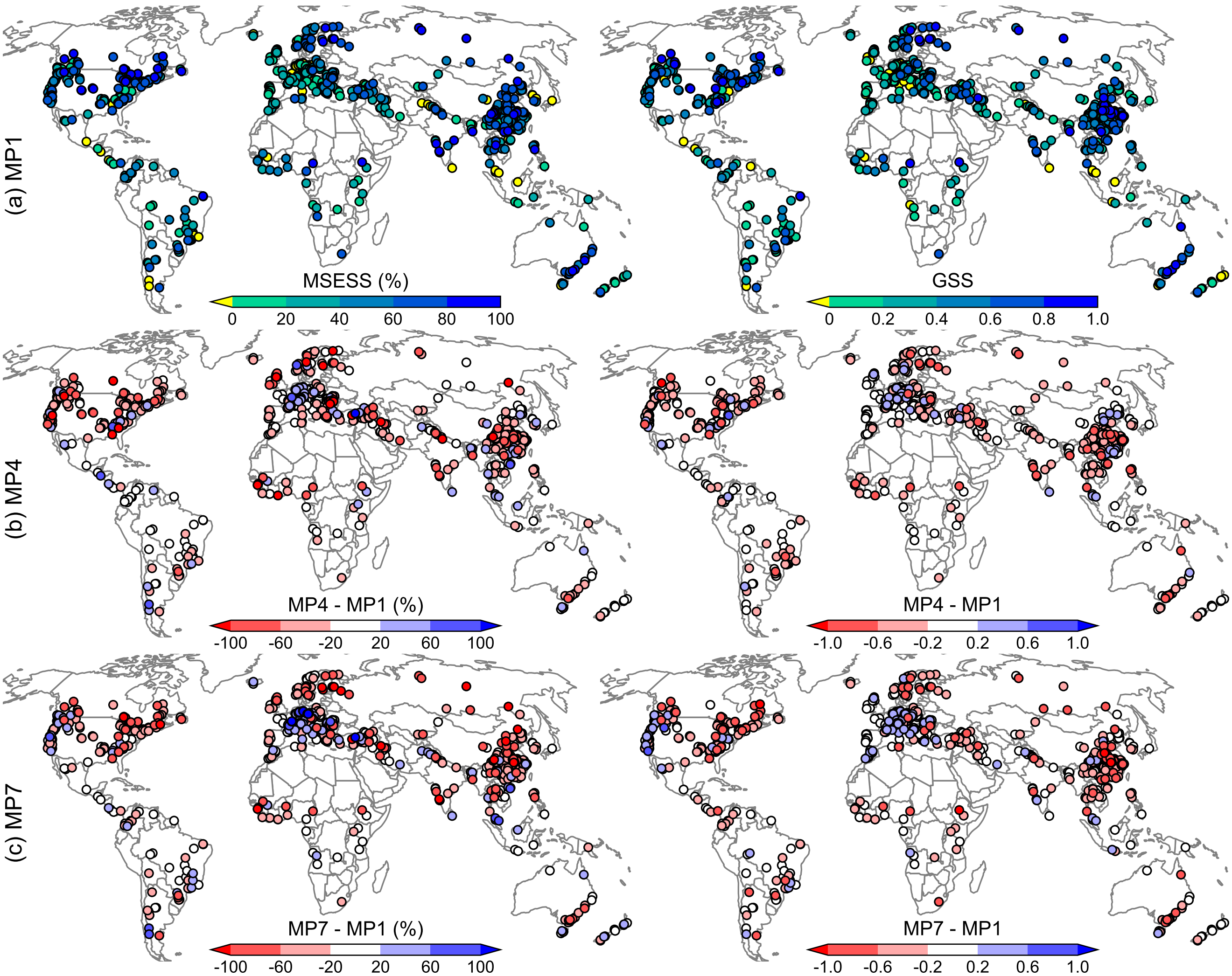

Figure S2. MSESS (left) and GSS (right) values for 735 dams. Taking a model with a lead-time of 1 month (MP1) as reference (a), we report the difference between MP1 and MP4 (b) and MP1 and MP7 (c).

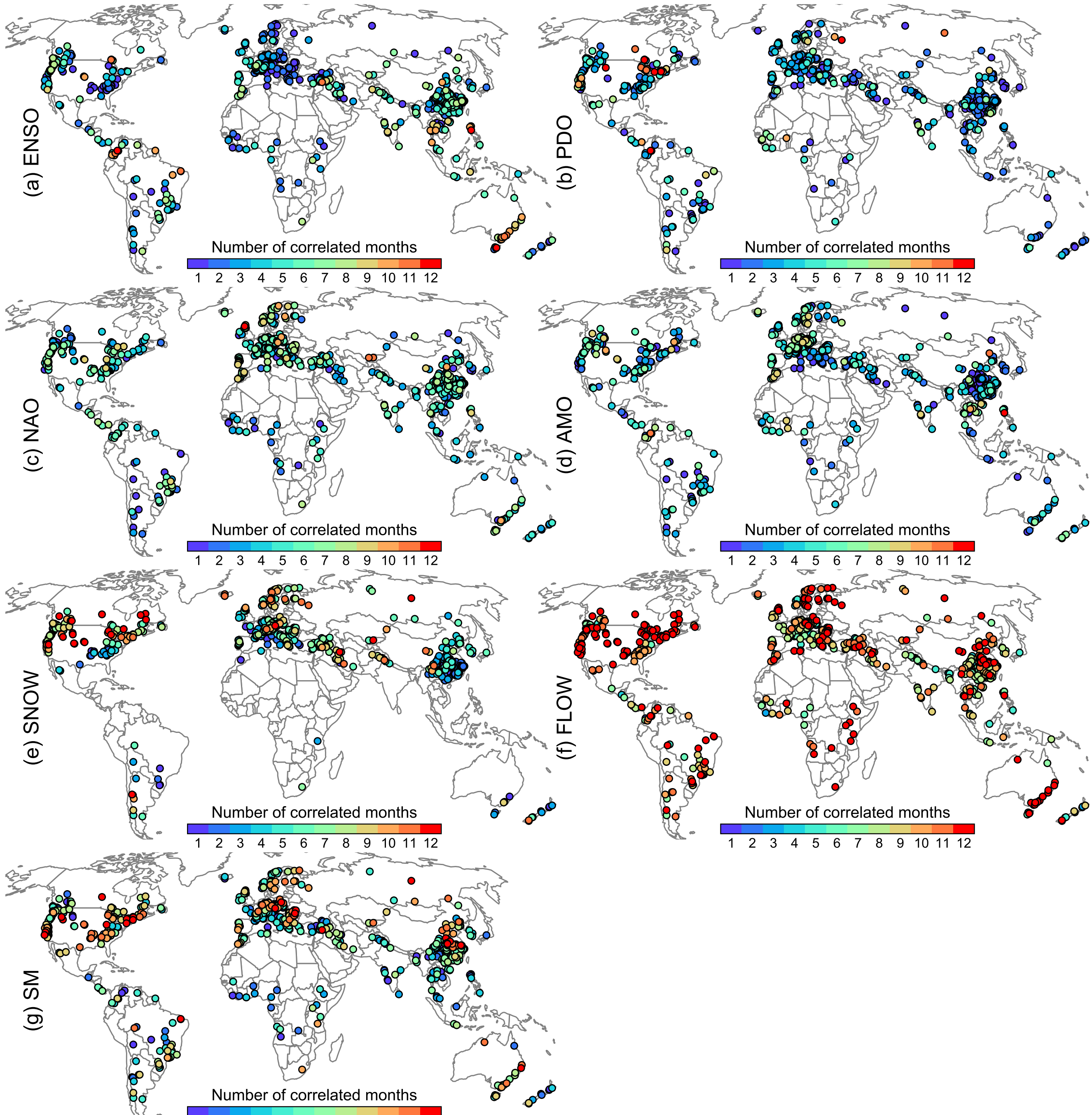

Figure S3. Number of months in which the monthly inflow is correlated with the lagged (a) ENSO, (b) PDO, (c) NAO, (d) AMO, and (e) snowfall, and 1 month ahead (f) inflow and (g) soil moisture drivers.

Table S1. Correlation between dam design specifications and performance gain I_{PF}.

No.	Variable	Description	Correlation with I_{PF}
1	x_{head}	Maximum hydraulic head	-0.332
2	x_{depth}	Ratio of max. reservoir depth to max. hydraulic head	0.362
3	x_{diff}	Max. hydraulic head - max. reservoir depth	-0.348
4	$\log(x_{empty})$	Ratio of storage capacity to max. turbine release	-0.449
5	$\log(x_{fill})$	Ratio of storage capacity to mean monthly inflow	-0.613

Table S2. Cross-validation scores of alternative formulations of the logistic regression model. The model is used to predict whether the performance gain I_{PF} is larger than 4.7% (the mean value of I_{PF} across the 735 dams).

Model			Accuracy	Kappa	Note
x_{depth}	+	x_{fill}	0.785	0.535	All factors are significant (*p*<0.01)
x_{depth}	+	$\log(x_{fill})$	0.776	0.503	All factors are significant (*p*<0.01)
x_{depth}	+	x_{empty}	0.766	0.488	All factors are significant (*p*<0.01)
x_{empty}	+	x_{fill}	0.750	0.468	x_{empty} is not significant (*p*=0.502)

Table S3. Sensitivity of logistic regression results to changes in the threshold used to divide dams into *success* ($I_{PF} >$ threshold) and *failure* ($I_{PF} \leq$ threshold) groups. Different formulations were tested (as in Table S2), but only models with accuracy higher than the default model are reported.

Threshold	$Success$	$Failure$	Model			Accuracy	Kappa
3%	56.2%	43.8.4%	x_{depth}	+	x_{fill}	0.814	0.617
4.7% (mean I_{PF})	36.6%	63.4%	x_{depth}	+	x_{fill}	0.785	0.535
7%	19.5%	80.5%	x_{depth}	+	x_{fill}	0.806	0.306
7%	19.5%	80.5%	x_{depth}	+	x_{empty}	0.821	0.353

Table S4. Correlation between dam design specifications (or forecast skill) and I.

No.	Variable	Description	Correlation with I
1	x_{head}	Maximum hydraulic head	-0.361
2	$\log(x_{fill})$	Ratio of storage capacity to mean monthly inflow	-0.295
3	x_{exceed}	Fraction of time inflow $>$ max. turbine release	0.355
4	x_{exceed}^{con}	Longest consec. months that inflow $>$ max. turbine release	0.355
5	x_{high}	Longest consec. months that inflow $\geq$ mean inflow	0.349
6	x_{hurst}	Hurst coefficient of annual inflow	0.262
7	x_{acf}	Lag 1 autocorrelation of annual inflow	0.287
8	x_{MARE}	Mean absolute relative error	-0.315
9	x_{MdAPE}	Median absolute percentage error	-0.400
10	x_{MRE}	Mean relative error	-0.257
11	x_{RSqr}	Coefficient of determination	0.251
12	x_{NSE}	Nash-Sutcliffe efficiency	0.252
13	x_{MSLE}	Mean squared logarithmic error	-0.306
14	x_{VE}	Volumetric efficiency	0.337

Table S5. Linear regression models for predicting I using 2–5 explanatory variables.

Model	Adj R-squared	Note
$x_{MdAPE} + x_{exceed}$	0.310	All factors are significant (p<0.01)
$x_{MdAPE} + x_{exceed} + x_{high}$	0.356	All factors are significant (p<0.01)
$x_{MdAPE} + x_{exceed} + x_{high} + x_{head}$	0.392	All factors are significant (p<0.01)
$x_{MdAPE} + x_{exceed} + x_{high} + x_{head} + x_{hurst}$	0.407	All factors are significant (p<0.01)

References

Alcamo, J., Döll, P., Henrichs, T., Kaspar, F., Lehner, B., Rösch, T., and Siebert, S.: Development and testing of the WaterGAP 2 global model of water use and availability, Hydrological Sciences Journal, 48, 317–337, 2003.

Dawson, C. W., Abrahart, R. J., and See, L. M.: HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts, Environmental Modelling & Software, 22, 1034–1052, 2007.

Kaveh, K., Hosseinjanzadeh, H., and Hosseini, K.: A new equation for calculation of reservoir's area-capacity curves, KSCE Journal of Civil Engineering, 17, 1149–1156, 2013.

Nowak, K., Prairie, J., Rajagopalan, B., and Lall, U.: A nonparametric stochastic approach for multisite disaggregation of annual to daily streamflow, Water Resources Research, 46, 2010.

Vu, D., Dang, T., Galelli, S., and Hossain, F.: Satellite observations reveal thirteen years of reservoir filling strategies, operating rules, and hydrological alterations in the Upper Mekong River Basin, Hydrology and Earth System Sciences Discussions, 2021, 1–28, https://doi.org/10.5194/hess-2021-360, 2021.

Weedon, G., Gomes, S., Viterbo, P., Shuttleworth, W. J., Blyth, E., Österle, H., Adam, J., Bellouin, N., Boucher, O., and Best, M.: Creation of the WATCH forcing data and its use to assess global and regional reference crop evaporation over land during the twentieth century, Journal of Hydrometeorology, 12, 823–848, 2011.

Wilks, D. S.: Statistical methods in the atmospheric sciences, vol. 100, Academic press, 2011.

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., and Pavelsky, T. M.: MERIT Hydro: A high-resolution global hydrography map based on latest topography dataset, Water Resources Research, 55, 5053–5073, 2019.