

2026-10-01-114355 Notes on Assessing trends in irrigation water withdrawal using a data-driven temporal extrapolation framework (Zarpas et al., 2026), Preprint

Abstract

🔗 The scientific problem: Assessing long-term trends in irrigation water withdrawal (IWW) under observational scarcity ([Zarpas et al., 2026, p. 1](#))

“Assessing long-term trends in irrigation water withdrawal (IWW) is essential for understanding the impacts of climate change on water resources and agricultural systems, yet long-term observations remain scarce.” ([pdf](#))

🔗 France case study: longitudinal availability ([Zarpas et al., 2026, p. 1](#))

“In France, annual IWW observations are only available from 2008 onward through the national water withdrawal database, limiting the analysis of long-term changes.” ([pdf](#))

🔗 Response: a data-driven extrapolation framework based on generalized additive models ([Zarpas et al., 2026, p. 1](#))

“To address this limitation, we develop a data-driven framework to reconstruct annual catchment-scale IWW over the 2000–2022 period. The approach relies on a generalized additive model (GAM), using structural (e.g., area equipped for irrigation) and hydro-meteorological (e.g., monthly mean temperature) predictors.” ([pdf](#))

🔗 IWW Decomposition strategy via GAMs ([Zarpas et al., 2026, p. 1](#))

“The model explicitly identifies inter-catchment and inter-annual variability through a decomposition of total IWW into a base IWW term and a relative deviation term. The model is evaluated using a series of Leave-One-Period-Out cross-validation experiments. Results show robust predictive performance across all experiments ($RMSE < 2$ mm/yr) and reliable uncertainty estimates.” ([pdf](#))

🔗 GAM Reconstruction on the 2000-2022 period (Extrapolation 2000-2007) and trend analysis ([Zarpas et al., 2026, p. 1](#))

“The framework was then used to reconstruct annual IWW over the 2000–2007 period and to generate continuous time series of IWW at the catchment scale over 2000–2022. The reconstructed time series reveal widespread positive trends, with significant increases detected in 92% of catchments. Trend analysis indicates that changes in total IWW are primarily driven by the relative deviation component. A regional synthesis further identifies the Mediterranean region as a major hotspot, combining high IWW and strong IWW positive trends.” ([pdf](#))

Introduction

🔗 Human influences on the water cycle ([Zarpas et al., 2026, p. 1](#))

“The explicit representation of human influences on the water cycle in hydrological models is essential for assessing past, current and future water availability.” ([pdf](#))

🔗 Goal of identifying human influences: climate change impact assessment and hydrological modeling ([Zarpas et al., 2026, p. 1](#))

“This is required in particular for climate change impact assessment, in order to efficiently manage river basins and to inform climate adaptation public policies (Galelli et al., 2025; Kreibich et al., 2025). To integrate these influences into a hydrological modelling framework, the different uses of water have to be identified, described in terms of their temporal and spatial variabilities, and quantified.” ([pdf](#))

🔗 Difficulty in obtaining observational (e.g.: survey) data ([Zarpas et al., 2026, p. 1](#))

“This often requires a substantial amount of field surveys, data collection and post-processing, which is a challenge even at the local scale (Santos et al., 2026; Vu et al., 2020).” ([pdf](#))

🔗 Irrigated agriculture as a key water use ([Zarpas et al., 2026, p. 1](#))

“Irrigated agriculture is a major water use that can significantly affect water availability (Puy et al., 2025; Yu et al., 2025), especially when its water demand peaks during summer low-flow periods (Mehta et al., 2024; Nie et al., 2021).” ([pdf](#))

☞ Challenges in IWW estimation: local knowledge fragmentation / lack of long-term consolidated data sources ([Zarpas et al., 2026, p. 1-2](#))

“Accurately estimating irrigation water withdrawal (IWW) volumes at large scales or within large-sample hydrological frameworks, however, presents several challenges (Puy et al., 2022b). Among them, we can cite the fragmented knowledge on actual local water use (e.g. irrigation practices, farmers decision calendars for irrigation, crop water requirements, artificial water storages), and the lack of long-term observational databases containing reliable information on where, when and how much water is withdrawn for irrigation to support IWW estimations and uncertainty quantification (Galelli et al., 2025). Accurate and temporally consistent estimates are nevertheless essential to support robust hydrological modeling (Nicolle et al., 2021; Thirel et al., 2015), and in general, to assess irrigation pressure on river system under changing climate and socio-economic conditions (Arboleda-Obando et al., 2025; Citrini et al., 2025; Tian et al., 2023).” ([pdf](#))

☞ Modeling as an alternative to complete observational data sources ([Zarpas et al., 2026, p. 2](#))

“In this context, modelling offers a valuable compromise for obtaining a coherent, complete, and continuous representation of irrigation dynamics across space and time (Linés and Werner, 2025; Portoghesi et al., 2025). It allows to leverage existing knowledge to fill spatial and temporal data gaps where direct observations are missing (Dang et al., 2024; Gorguner and Kavvas, 2020).” ([pdf](#))

☞ IWW Modeling via process-based and statistical (data-driven) models ([Zarpas et al., 2026, p. 2](#))

“Modelling IWW is usually done through process-based or statistical models, and, in itself, presents its own challenges. Process-based models explicitly represent known energy and water fluxes as well as human decision-making processes (De Wit et al., 2019; Gaudou et al., 2014), but they present limitations as they often rely on poorly constrained parameters (Linga et al., 2025; Soutif-Bellenger et al., 2023; Puy et al., 2022a). In recent years, this limitation has stimulated the development of data-driven approaches for IWW estimation (Kragh et al., 2025; Zhang et al., 2025).” ([pdf](#))

📖 Excerpt ([Zarpas et al., 2026, p. 2](#))

“In particular, the Random Forest (RF) algorithm, has shown strong performance in interpolation tasks, although it exhibits known structural limitations when used for extrapolation, i.e., when predicting beyond the conditions of its training sets (Muckley et al., 2023; Von Korff and Sander, 2022). This limitation has been highlighted by the difficulty of capturing interannual variability in IWW for a given catchment when temporal information is excluded from model training (e.g. a specific year) (Zarpas et al., 2026b). This can be explained by the fact that RF does not explicitly account for the spatiotemporal autocorrelation structure of IWW (Zarpas et al., 2026b). **Methodological developments are therefore needed to enable reliable temporal extrapolation of IWW by data-driven models.** The relationship between IWW and its main predictors can be reasonably approximated by an additive structure (Zarpas et al., 2026b; Zhang et al., 2025). Under this assumption, additive models, such as the semi-parametric Generalized Additive Models (GAM, Hastie and Tibshirani, 1986), represent a promising way to perform extrapolation, as they mitigate the curse of dimensionality. GAMs provide smooth and continuous functional relationships and can naturally accommodate structured spatiotemporal effects. This is essential because data-driven IWW estimation from short reference time series, available on multiple sites, can be viewed as a spatiotemporal problem (Shortridge, 2023; Das et al., 2018), and particularly in the context of temporal extrapolation.” ([pdf](#))

✖ Issue

"Methodological developments are therefore needed to enable reliable temporal extrapolation of IWW by data-driven models."

This is an unjustified claim when the only reference data-driven method presented is the RF, which is but only one of many possible ML / data-driven approaches, many of which are suited for time-series prediction and can explicitly model autocorrelation structure (even the RF can, depending on the specific design of the training and inference paradigm).

📖 Goal of the Study ([Zarpas et al., 2026, p. 2](#))

“The objective of this study is to develop and evaluate a GAM based data-driven framework for the temporal extrapolation of annual IWW at the catchment scale, and to exploit the reconstructed time series to improve our understanding of long-term IWW dynamics.” ([pdf](#))

🔗 Instrumental use of the extrapolated IWW series for analysis ([Zarpas et al., 2026, p. 2](#))

“The resulting temporally continuous IWW series enables subsequent time-series analyses, with trend assessment considered here as an illustrative application.” ([pdf](#))

× Major Issues

1. **The conclusions derived from this study are entirely based on the model.** With only AEI and the meteorological anomalies varying over time, the detected trends are fully determined by trends in those predictors. Changes in efficiency, crops, regulation or BNPE reporting *simply cannot* appear. This also makes "trends are driven mainly by dW" true largely by construction (Table 1 assigns all climate signal to dW).
2. **In this study, there is no benchmarking against established baselines (e.g.: climatology, per-catchment climatology).** I think this is a liability for the study, since what you wish to demonstrate is extrapolation capability / explanation of year-to-year variability, although your metrics do not clearly isolate that from just being good, on average, across the catchments. Indeed, you compute metrics that are pooled across the catchments, yet you do not show how much the between-catchment variability dominates over the within-catchment's: a Coefficient of determination R^2 can only tell you how much better your predictions are compared to predicting with the *global mean*, but a more informative baseline would be the per-catchment mean, which your model can happily achieve just by adjusting the Z_c 's. Also, see my comments below about your choice of metrics.

🔗 (Towards) Method: assessing extrapolation capabilities of the GAM ([Zarpas et al., 2026, p. 2](#))

“To this end, we first assess the temporal extrapolation capability of the proposed GAM, using a demanding temporal cross-validation framework, while also quantifying prediction uncertainty. We then analyze IWW trends at the catchment scale.” ([pdf](#))

🔗 **Scope: Catchment-scale analysis over 498 sites in France** ([Zarpas et al., 2026, p. 3](#))

“The methodology is applied to the French national water withdrawal database (BNPE), which provides annual data from 2008 to 2022. It allows us to reconstruct continuous IWW time series over the 2000–2007 period. To produce IWW estimates, suitable for representing irrigated agriculture as a major anthropogenic influence on the water cycle, the analysis is carried out at the catchment scale over a dataset covering 498 catchments in France.” ([pdf](#))

Method

🔗 **Excerpt** ([Zarpas et al., 2026, p. 3](#))

“We propose a data-driven framework to extend in time a reference annual IWW database.” ([pdf](#))

× **Presentation issue**

Unnecessary repetition

🔗 **Sources of variability** ([Zarpas et al., 2026, p. 3](#))

“Because the reference database spans numerous catchments and multiple years, two distinct sources of variability emerge: the inter-catchment variability and the intra-catchment inter-annual variability.” ([pdf](#))

🔗 **Presentation**

Good way to frame the problem, could benefit from being moved further up so that the reader can be exposed to this guiding principle up front

🔗 **Excerpt** ([Zarpas et al., 2026, p. 3](#))

“Building on the decomposition proposed by Zarpas et al. (2026b),” ([pdf](#))

× **Issue**

I am unsure as to how much the methodology of this paper is innovative compared to Zarpas et al. (2026b); I would like the authors to clarify exactly what is new relative to (2026b). Also: the GAM is motivated by RF failing to “account for the spatiotemporal autocorrelation structure”, yet the GAM has no temporal correlation term either. Its errors are i.i.d. and Z_c is a constant per catchment.

§§ Decomposition of the total IWW ([Zarpas et al., 2026, p. 3](#))

“[...] we decompose the total IWW (denoted as $\mathcal{W}_{y,c}$ for catchment c and year y) into 1) a structural component referred to as the base IWW term (denoted $\mathcal{W}_{y,c}^0$), representing the background level of IWW associated with long-term catchment characteristics and reference climatological conditions, while evolving over time through slowly varying anthropogenic predictors, and 2) an interannual component referred to as the relative deviation term (denoted $\delta\mathcal{W}_{y,c}$), describing year-to-year fluctuations from the structural level:

$$\mathcal{W}_{y,c} = \mathcal{W}_{y,c}^0(1 + \delta\mathcal{W}_{y,c}). \quad (1)$$

” ([pdf](#))

§§ Hierarchical structure of nested catchments in the GAM formulation ([Zarpas et al., 2026, p. 3](#))

“The spatio-temporal organization of the reference database has important implications for the GAM model formulation, as it induces a hierarchical dependence structure. In particular, some catchments are nested within others, inducing spatial correlation through shared drainage areas. The GAM formulation presented below is specifically designed to account for this hierarchical spatial dependence.” ([pdf](#))

Logarithmic-scale modeling of the IWW per year & catchment (Zarpas et al., 2026, p. 3)

“We model IWW in the logarithmic space to ensure strictly positive predictions of IWW, and to enable the identification of both the additive GAM decomposition and the multiplicative IWW variability decomposition of $\mathcal{W}_{y,c}$. We fit a GAM (f_{GAM}), using R package mgcv (Wood, 2011), as specified in Equation (2):

$$\log(\mathcal{W}_{y,c}) = f_{GAM}(\mathbf{X}_{y,c}) + \varepsilon_{y,c} = \sum_{i=1,\dots,n} f_i(x_{i,y,c}) + Z_c + \varepsilon_{y,c}, \quad (2)$$

where $\mathbf{X}_{y,c}$ is the predictor vector of individual component $x_{i,y,c}$; f_i is the individual smoother of the i -th component of the predictors, and n being the number of components of $\mathbf{X}_{y,c}$ and the summation running over all individual predictors; Z_c is a random effect - which introduces a hierarchical structure into the model (Pedersen et al., 2019); and $\varepsilon_{y,c}$ is the error, for the year y and the catchment c . Spatial correlation is accounted for in the construction of the random effect (see Appendices A and B for details).” (pdf)

Definition of the "smoothed" components (Zarpas et al., 2026, p. 24)

“Our GAM model can be written in a simplified form as:

$$f_{GAM}(\mathbf{X}_{y,c}) = \sum_{i=1,\dots,n} f_i(x_{i,y,c}) + Z_c = \beta \mathbf{B}_{y,c} + Z_c, \quad (A1)$$

where $\mathbf{B}_{y,c}$ denotes the vector of spline basis functions evaluated at the predictor values, and β the corresponding fitted coefficients.” (pdf)

⚠ Clarity issue

The fact that the f_i represent a spline-transform of the native feature i should be mentioned up front in the methodological section and not be relegated to an appendix. Also, spline smooths extrapolate linearly, which might not be ideal for temporal extrapolation. This limitation should be stated explicitly.

§§ Theoretical derivation of the correlation in the spatial random effect from Appendix B ([Zarpas et al., 2026, p. 25](#))

“To derive a theoretical expression of the correlation in the spatial random effect, we assume that within each catchment the IWW is randomly distributed with expectation μ and variance V . Let us further assume that it is independently and identically distributed across all catchments. In this framework, the correlation between two disjoint catchments is null and the correlation between nested catchment is given by how much the smallest catchment is nested in the largest one. More formally, let us consider two catchments C_i and C_j , with $i \neq j$, and let us denote A_i and A_j their respective areas. Without loss of generality, we suppose $C_i \subset C_j$, hence $A_i \leq A_j$. We further denote W_i and W_j the random IWW on C_i and C_j , respectively. With these assumptions,

$$\text{Cov}(W_i, W_j) = \frac{1}{A_i A_j} \int_{C_i} \int_{C_i} \text{Var}(W_i) d\mathbf{u} d\mathbf{v} + \frac{1}{A_i A_j} \int_{C_i} \int_{C_j \setminus C_i} \text{Cov}(W_i, W_j) d\mathbf{u} d\mathbf{v}$$

with $A_i/A_j < 1$. Since $\text{Var}(W_i) = \text{Var}(W_j) = V$, we thus get that the correlation is $\rho(W_i, W_j) = A_i/A_j < 1$. Finally, one thus gets in all generality

$$\rho(W_i, W_j) = \min(A_i, A_j) / \max(A_i, A_j). \quad (\text{B2})$$

” ([pdf](#))

× Issues

1. Aside from the fact that, as it is written, this paragraph is contradictory (the claim that catchments' IWW are IID is contradicted right after since independent $\Rightarrow \text{Cov} = 0$, whereas nested catchments are in fact spatially correlated), it does not make sense for two catchments of different areas to have the same variance, since no additive model offers that. Assuming that the total withdrawal volumes are proportional to the catchment areas, e.g., akin a Poisson point process representing random unit-volume withdrawals, then the catchment-area averages (in mm/yr) would have variances *inversely proportional to areas* (see point 4.), but the independence of disjoint catchments and a similar covariance structure could still be achieved (see below).
2. Tightly-linked to the previous: you conflate the natural and logarithmic scales for the IWW. In the method section, the error is additive with respect to $\log(\mathcal{W}_{y,c})$ which is by your own definition the *logarithm* of the IWW in catchment c at year y ; in this appendix, you provide a statement for the IWWs over different catchments in *natural* scale, i.e., W_c (also, what is the reason for the change in notation $\mathcal{W} \rightarrow W$?).
3. Further: you do not make it clear how the correlation in natural scale between IWW can be used to constrain / influence the correlation of the errors $\varepsilon_{y,c}$ in the log scale, although you do claim that "*Spatial correlation is accounted for in the construction of the random effect (see Appendices A and B for details).*"
4. Equation B1 is nonsensical as written: the same covariance appears on both sides (once inside and once outside the integral), and each W_i was never individually defined.

I take it you mean to define each W_i , e.g., as the integral of a random field over catchment C_i , over the corresponding area A_i i.e.:

$$W_i := \frac{1}{A_i} T_i, \quad T_i := \int_{C_i} w(\mathbf{u}) d\mathbf{u},$$

with w a random field defined over the entire domain, and fulfilling independence under disjoint catchments, such as a Poisson point process as suggested in point 1. Notice that independence is a very strong assumption on the IWW's random field w , which is unlikely to hold in practice (e.g., neighboring catchments will tend to have correlated IWWs). Also, notice that the T_i (total withdrawal volumes of catchments) have a mean and variance proportional to the areas A_i , which means that the transformations W_i have variances inversely proportional to areas.

5. Net of the specific definition of the W_i over each catchment's area, a much more straightforward justification of the monotonicity of the covariance $\text{Cov}(W_i, W_j)$ wrt. the area ratio A_i/A_j for nested catchments

$C_i \subset C_j \Rightarrow A_i < A_j$ could (and perhaps should) have been provided: assuming

- additivity of the total volume withdrawals $T_j = T_i + T_{j-i}$ (e.g., trivial under the first example definition of point 2),
- independence of disjoint catchments' IWWs ($C_u \cap C_v = \emptyset \Rightarrow T_u \perp T_v$), and
- that variance of each T_i is proportional to the area of its catchment (reasonable if one assumes that "unit" withdrawals are a stationary point process on the map, e.g., in a Poisson point process such as the one in point 4), and consequently the variance of the IWWs as defined in the paper (over catchment areas, i.e., in mm/yr) are *inversely proportional to areas*

then

$$\text{Cov}(T_i, T_j) = \text{Cov}(T_i, T_i + T_{j-i}) = \text{Var}(T_i) + \text{Cov}(T_i, T_{j-i}) = \text{Var}(T_i)$$

and as a consequence

$$\rho(W_i, W_j) = \rho(T_i, T_j) = \frac{\text{Var}(W_i)}{\text{Var}(W_i) \sqrt{\text{Var}(W_j)/\text{Var}(W_i)}} = \sqrt{\frac{A_i}{A_j}} < 1.$$

6. Equation B2 is wrong unless the two catchments are one nested in the other; the corrected general statement under 3. (valid even when catchments are disjoint) is

$$\rho(W_i, W_j) = \frac{|C_i \cap C_j|}{\sqrt{A_i A_j}}.$$

7. In general, you should probably justify the choice of having non-independent units a bit more in detail, e.g., as a consequence of having non-overlapping gaps in feature or targets over nested catchments (even just a sentence along these lines would greatly help). Otherwise, it sounds like you could have simply defined the total withdrawal volumes to be independent by subtraction (e.g.: in the examples above, the new units would become W_i, W_{j-i}), which however require some form of attribution of features to the newly-created unit $j - i := C_j/C_i$. If this is the limitation you face, please state so, because imposing a given covariance structure in a GAM is not the same as requiring the additivity property.

“We note that in the log-space the IWW decomposition becomes:

$$\log(\mathcal{W}_{y,c}) = \log(\mathcal{W}_{y,c}^0) + \log(1 + \delta\mathcal{W}_{y,c}). \quad (4)$$

The identification between IWW components and the GAM decomposition thus provides the following estimators:

$$\begin{aligned} \log(\hat{W}_{y,c}^0) &= \sum_{i=1, \dots, n^0} f_i^0(x_{i,y,c}^0) + Z_c \\ \log(1 + \delta\hat{W}_{y,c}) &= \sum_{i=1, \dots, n^\delta} f_i^\delta(x_{i,y,c}^\delta) \end{aligned} \quad (5)$$

” ([pdf](#))

⚠ Issues

3. Concerning notation: you have changed the symbols ($\mathcal{W} \rightarrow W$) and the "hat" should sit outside of the logarithm on both subequations
4. I think you should have introduced this decomposition as a *design choice* rather than exposing it here as if it was something that follows equations (2) or (3), since that decomposition is not unique: any pair of affine combinations

$$\begin{cases} L_1 &= \alpha \log(\hat{W}_{y,c}^0) + (1 - \alpha) \log(1 + \delta\hat{W}_{y,c}) \\ L_2 &= (1 - \alpha) \log(\hat{W}_{y,c}^0) + \alpha \log(1 + \delta\hat{W}_{y,c}) \end{cases}$$

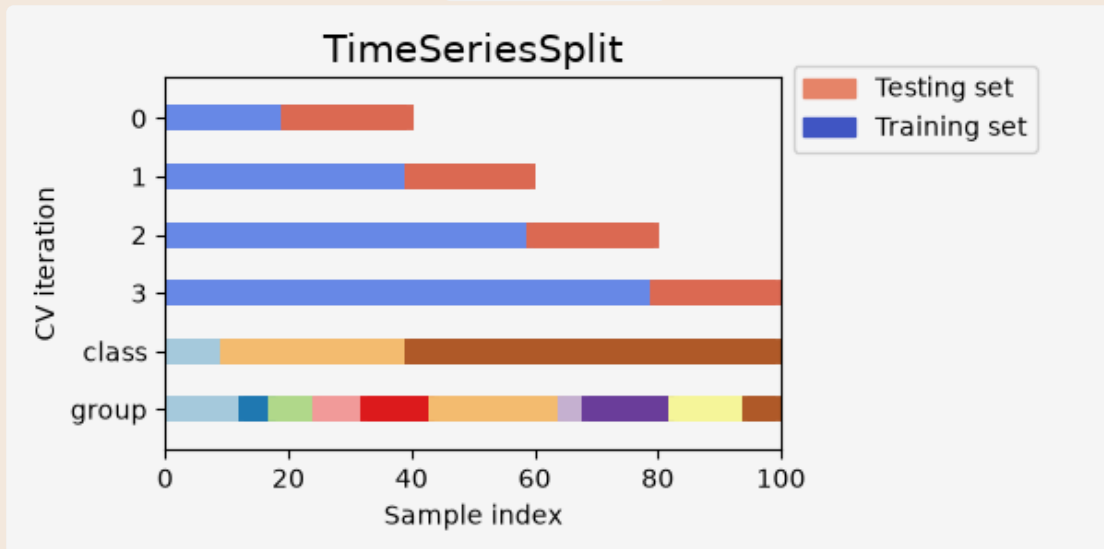
for $\alpha \in \mathbb{R}$ fulfills the required property. The issue is that from how you have set up the problem so far $\mathcal{W}_{y,c}^0$ and $\delta\mathcal{W}_{y,c}$ are *not identifiable* so it's better to say it early on and explicit the fact that you attach meaning to the two based on a *definition* such as equation (5).

Blocked Cross-Validation via "Leave One Period Out" (LOPO) ([Zarpas et al., 2026, p. 4](#))

“To assess the ability of our model to produce reliable temporal extrapolation (here, on the 2000–2007 period) at the catchment scale, we perform temporal cross-validation experiments (Roberts et al., 2017) on the reference period (here, 2008–2022), where observations are available. Each experiment follows a Leave One Period Out (LOPO) strategy, in which the reference period is partitioned into blocks of consecutive years so that each block is in turn used as a test set for validation, while the remaining blocks form the training set. The length and, consequently, the number of time blocks used for validation varies from one LOPO experiment to another. We performed five experiments with 2, 3, 5, and 15 partitions of the reference period, respectively.” ([pdf](#))

? Question

I would consider, if time allows, using a different blocking strategy, similar to the next figure (taken from the `scikit-learn` documentation)^[1]



albeit in an opposite direction, since your goal is to reconstruct the past.

The reason this might be more robust is that your model might be flattered by "pinning" the endpoints of the extrapolation window, which happens on most test folds.

Interior blocks (most blocks in LOPO3, LOPO5 and LOPO15) are temporal interpolation. Roberts et al. (2017), which you also cite, also recommends buffer years between blocks because of autocorrelation.

📖 Excerpt ([Zarpas et al., 2026, p. 4](#))

“It is to be noted that none of these experiments is a clear reference for our extrapolation over the 2000-2007 period (after training over the entire 2008-2022 period).” ([pdf](#))

 Comment

The authors frame this "lack of reference" as a consequence of no extrapolation experiment simultaneously having a comparable-length training and test period, although I'd ascribe it to what is described in my comment to the LOPO strategy

🔗 Evaluation Protocol ([Zarpas et al., 2026, p. 5](#))

“The overall model performance is evaluated by considering all catchments pooled together, and by using the Root Mean Squared Error (RMSE) and the squared Pearson correlation coefficient r^2 . Model performance is also evaluated at the catchment scale (i.e., on each catchment individually), using the predictive error normalized by the observed catchment mean. This Catchment Relative Error (CRE) is defined for a variable v estimated by $\hat{v}$ for a year y and catchment c as:

$$\text{CRE}_{y,c} = \frac{v_{y,c} - \hat{v}_{y,c}}{\bar{v}_{y,c}} \quad (6)$$

” ([pdf](#))

× Issues

5. [minor] Pooling observations together might artificially inflate the importance of the most data-rich catchments and deflates the most data-poor, which I find problematic, especially since quantitative statements about missing data (e.g.: counts, per catchment) do not appear in the work. A better alternative would be, e.g., to average scores over catchments and accompany the metrics with bootstrapped intervals (sampling on catchments) and a sensitivity analysis based, e.g., on weighted re-sampling by a) catchment size b) historical average of IWW over catchment.
6. [major] Tightly-linked to the previous: as mentioned above, the issue with even the "correct" Coefficient of determination R^2 is that it naturally measures predictive performance relative to the *global mean*, which your model can easily beat just by adjusting the Z_c . At least some indication of intra-catchment variability should be provided, if not a full per-catchment performance skill-against-baseline assessment.
7. [major] It looks like you are using the square of Pearson's Correlation coefficient (r), which is defined as

$$r = \frac{\text{Cov}(Y, \hat{Y})}{\sigma_Y, \sigma_{\hat{Y}}} \approx \frac{\sum_{i=1}^n (y_i - \bar{y}) (\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}}, \quad (\text{r1})$$

with the Coefficient of Determination (R^2), defined as

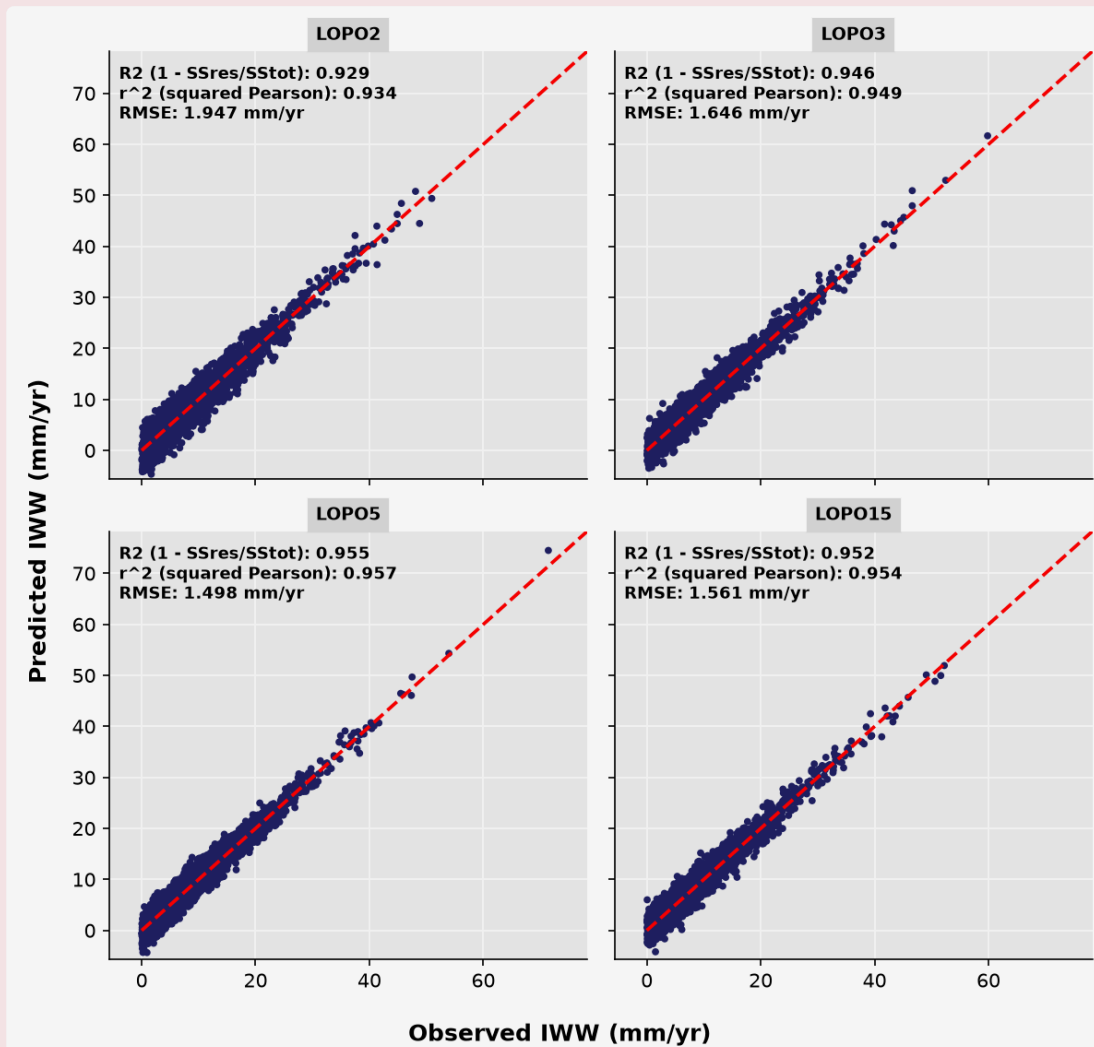
$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{n \text{RMSE}(\mathbf{y}, \hat{\mathbf{y}})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \quad (\text{r3})$$

The R^2 is a valid metric performance metric in regression, since its maximum value (1) is achieved **only** when $y_i \equiv \hat{y}_i$, which is **not** the case for Pearson's r : its maximum value is 1 whenever the two series (y_i and $\hat{y}_i$) are *an increasing linear transformation of one another*, i.e., $y_i = a\hat{y}_i + b$, $a > 0$, which completely neglects bias and factors of scale affecting the predictions; predicting $\hat{y}_i$, $1000 \cdot \hat{y}_i$ and $\hat{y}_i - 1000000$ are scored identically under this metric; further, you *square* the correlation coefficient, which lands its *minimum* value of -1 back to 1, i.e., the maximum possible score is identically achieved by predicting y_i with $-y_i$.

I think this confusion is quite a big oversight for an academic work, if not an attempt to mix one known metric with another that is slightly more flattering.

To convince myself this wasn't just a typo, I tried to replicate the plots in Figure 2 with some synthetic data $y_i \sim \text{Gamma}(\alpha, \beta)$, $\hat{y}_i = y_i + e_i$, with e_i an

IID gaussian noise of mean 0 and variance σ , with α, β and σ chosen to match the coefficient of determination and RMSE reported for each LOPO variant, resulting in the following plots:



Visually the difference is apparent, and the squared Pearson's coefficient consistently inflates the performance compared to the correct R^2 . What I find even worse is that this plot assumes an *unbiased* prediction, i.e., $\mathbb{E}[\hat{y}_i] = y_i$, whereas the plots in the paper might mask a bias in the thick of the distribution, where dots are too dense to distinguish them individually.

I think the authors should provide at least the R^2 upon a major revision, if not NSE/KGE and a breakdown of its components.

§§ Uncertainty Modeling ([Zarpas et al., 2026, p. 5](#))

“Uncertainties are modeled using the heteroscedastic log-normal distribution (LN *) detailed in Appendix D1, i.e. we consider the estimator of the IWW $\hat{\mathcal{W}}_{y,c}$ as:

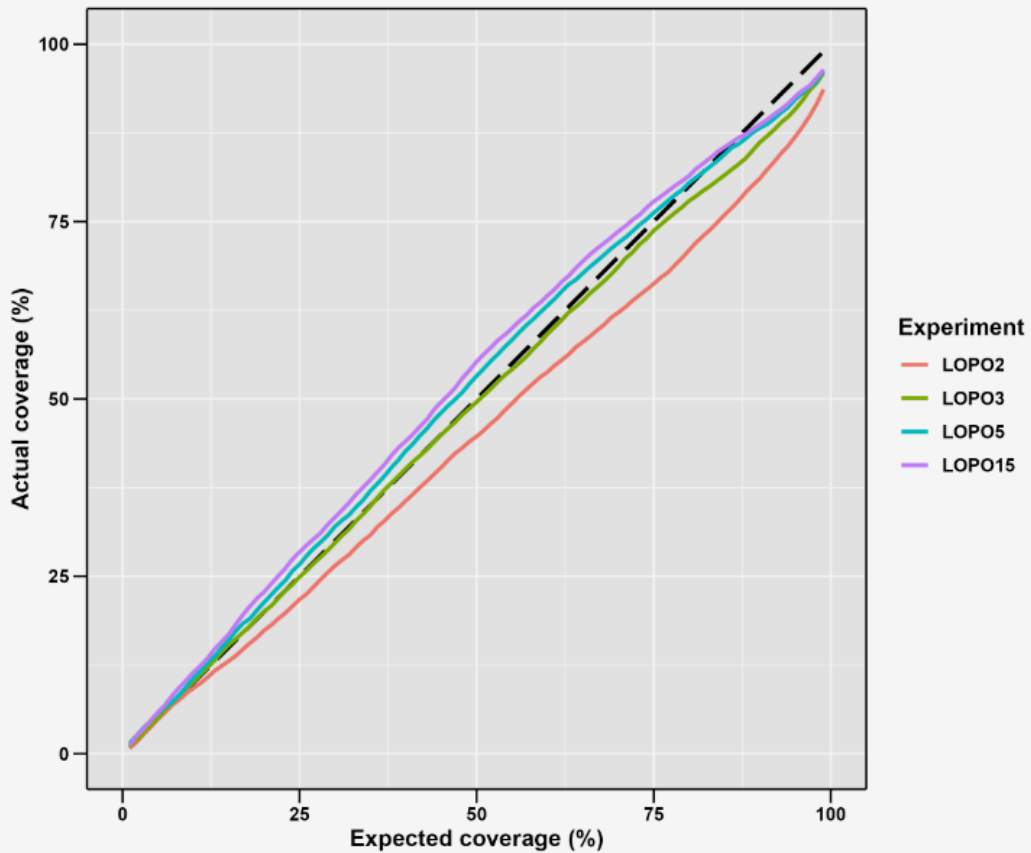
$$\hat{\mathcal{W}}_{y,c} | \mathbf{X}_{y,c} \sim \mathcal{LN}^*(f^{GAM}(\mathbf{X}_{y,c}), \sigma_{y,c}^2). \quad (7)$$

The log-scale variance ($\sigma_{y,c}^2$, whose derivation is detailed in Appendix D2 [...])” ([pdf](#))

§§ Empirical Verification of the Uncertainty Estimation ([Zarpas et al., 2026, p. 5](#))

“To assess the quality of this uncertainty estimation, we compare the nominal confidence level with the empirical coverage, i.e., the actual percentage of IWW data points contained within the confidence interval. This is performed for each crossvalidation experiment.” ([pdf](#))

§§ Figure 3 - Empirical justificaiton for the uncertainty estimate ([Zarpas et al., 2026, p. 13](#))



“**Figure 3.** Uncertainty estimation performance: actual coverage vs. expected coverage across confidence levels for all cross-validation experiments.” ([pdf](#))

§§ D1 "Modified" log normal law ([Zarpas et al., 2026, p. 26](#))

"We assume:

$$\log \hat{\mathcal{W}}_{y,c} | \mathbf{X}_{y,c} \sim \mathcal{N}(f^{GAM}(\mathbf{X}_{y,c}), \sigma_{y,c}^2) \quad (D1)$$

where $\sigma_{y,c}^2$ is a location-dependent variance term capturing heteroscedastic uncertainty.

Under this formulation, the induced distribution of $\mathcal{W}_{y,c}$ is log-normal. Here we use a modified log-normal distribution $\mathcal{LN}^*$ as:

$$\mathcal{LN}^* = \exp\left(-\frac{\sigma_{y,c}^2}{2}\right) \mathcal{LN} \quad (D2)$$

This transformation yields:

$$\mathbb{E}[\hat{\mathcal{W}}_{y,c}] = \exp(f^{GAM}(\mathbf{X}_{y,c})), \quad (D3)$$

" ([pdf](#))

§§ Excerpt ([Zarpas et al., 2026, p. 26](#))

"Our model is given by:

$$\log(\mathcal{W}_{y,c}) = f^{GAM}(\mathbf{X}_{y,c}) + \varepsilon_{y,c} \quad (D4)$$

We make the assumption that $\varepsilon_{y,c} \sim \mathcal{N}(0, \sigma_{\text{err}}^2)$ and accounts for model bias and irreducible noise. Neglecting the covariance between bias and variance of $f^{GAM}(\mathbf{X}_{y,c})$, the predictive uncertainty distribution is:

$$\log \hat{\mathcal{W}}_{y,c} | \mathbf{X}_{y,c} \sim \mathcal{N}\left(f^{GAM}(\mathbf{X}_{y,c}), \text{var}\left(f^{GAM}(\mathbf{X}_{y,c})\right) + \sigma_{\text{err}}^2\right) \quad (D5)$$

So, the total log-scale standard deviation is given by:

$$\sigma_{y,c} = \sqrt{\text{var}\left(f^{GAM}(\mathbf{X}_{y,c})\right) + \sigma_{\text{err}}^2}$$

" ([pdf](#))

× Major Issues

1. "Under this formulation, the induced distribution of $\mathcal{W}_{y,c}$ is log-normal.": the authors should state explicitly that they model separately the IWW and the *predictive distribution* for the IWW, which is **not the same as having a model for the estimator uncertainty**. What you define as $\hat{\mathcal{W}}_{y,c}$ should be clearly labeled as a predictive distribution and not as estimator uncertainty. From a statistical point of view, the properties of an estimator directly follow from assumptions about the model and how an estimator is constructed from data, **they are not a design choice**. I would **warmly** recommend using other, more rigorous ways to estimate model uncertainty, such as using bootstrap ensembles or applying asymptotic results (if any and the assumptions are met).
2. The **empirical justification** that they provide in support of the uncertainty estimate, i.e., comparing the nominal and observed coverage fractions (see Figure 3), **is by itself insufficient** to justify it, since even a completely nonsensical model for the uncertainty would look perfect in Figure 3: simply take the interval formed by the $1 - \alpha/2$, $1 + \alpha/2$ marginal quantiles over the observed y_i , regardless of covariates. The reason why this dummy model works here is that you do not measure the *calibration* of the coverage *conditionally* on the covariates; a better visual instrument would be a spread-skill plot (the predictions' variability on the x-axis, against the predictive performance on the y-axis), but only as a *diagnostics*, never as a justification for adopting a similar uncertainty model.
3. The same conditional density for $\hat{\mathcal{W}}_{y,c}|\mathbf{X}_{y,c}$ is given two different meanings across equations (7), (D1), (D5) (see further for an explanation).
4. "we use a modified log-normal distribution $\mathcal{LN}^*$ ": there is no such thing as a "modified" log-normal density in what you present, you neglecting a basic property of the log-normal distribution: let $W \sim \mathcal{LN}(\mu, \sigma^2)$, and $W' = aW$, $a > 0$, then

$$\log W' \sim \log a + \mathcal{N}(\mu, \sigma^2) = \mathcal{N}(\mu + \log a, \sigma^2) \Rightarrow W' \sim \mathcal{LN}(\mu + \log a, \sigma^2)$$

so what you have defined is just another lognormal distribution, with a log-location parameter offset by $\log a$, which in your case amounts to $-\sigma_{y,c}^2/2$. However, you gloss over the fact that you have implicitly set

$$\log \hat{\mathcal{W}}_{y,c}|\mathbf{X}_{y,c} \sim \mathcal{N}\left(f^{GAM}(\mathbf{X}_{y,c}) - \sigma_{y,c}^2/2, \sigma_{y,c}^2\right) \quad (\text{r4})$$

which now directly contradicts (D1) as well as (D5).

Also, the shift goes in the opposite direction from the usual retransformation correction, which multiplies by $\exp(+\sigma^2/2)$. The cited

references (Girard et al. 2004; Lütkepohl and Xu 2012) deal with that usual correction.

From what I can see here, the authors are trying to "model" the generalization properties of their estimators, but they confuse predictive distribution / uncertainty with model uncertainty. A clearer labeling of the type of uncertainty they are tackling and the literature in support of their method should be offered.

-
1. Image source: https://scikit-learn.org/stable/auto_examples/model_selection/plot_cv_indices.html#visualize-cross-validation-indices-for-many-cv-objects ↩