Response to Reviewer

Testing the Temporal and Spatial Transferability of a Water Balance Model

Degenne et al.

We thank Dr David Post for his positive assessment and for his careful reading of the manuscript. We are pleased that he recommends the paper for publication. Below we address each comment in turn.

General comment — Clarity on what is being estimated

It would be useful if the authors are clearer about exactly what is being estimated.

We agree that this deserves clarification. Our model estimates **annual streamflow** Q_n for each year *n* via Eq. (4), which combines two components:

1. A **long-term mean runoff** term: the Turc–Mezentsev formula (Eq. 3) provides a first estimate of mean interannual streamflow (MQ_{TM}) from the interannual means of precipitation and potential evapotranspiration. This estimate is then corrected by a multiplicative coefficient e_{MQ}, which is one of the four parameters of the model.

2. An **anomaly component**: three elasticity coefficients ($e_{Q/P}$, $e_{Q/E}$, $e_{Q/\Lambda}$) translate annual deviations in precipitation, potential evaporation, and their synchronicity into a streamflow anomaly around the mean.

The full model (Eq. 4) is therefore:

$$Q_n = e_{Q/P}\,\Delta P_n + e_{Q/E}\,\Delta E0_n + e_{Q/\Lambda}\,\Delta\Lambda_n + e_{MQ}MQ_{TM}$$

Depending on the estimation method, the four parameters are either learned globally by the neural network using catchment descriptors (hybrid approach) or estimated locally for each catchment independently by ordinary least squares (local regression approach). We will add a summary sentence early in the model description to make this structure immediately clear to the reader.

L52 — Why mention Australia specifically?

Why mention Australia specifically? Many other countries have a denser network of gauging stations.

Original text:

"In regions characterized by a dense network of gauging stations, such as in France or in Australia, approaches based on spatial proximity consistently outperform other methods (Merz and Blöschl, 2004; Oudin et al., 2008)."

The mention of Australia is motivated by the papers cited at this point (Merz and Blöschl, 2004; Oudin et al., 2008), which use France and Australia as canonical examples of regions where spatial proximity-based regionalization performs well due to dense gauge networks. The sentence is not intended to single out Australia as unique in this regard. We will rephrase it to make clear that France and Australia are cited here as examples used in those specific references, and not as the only countries with dense networks.

L180 — What is 'interannual memory'?

It is unclear to me what 'interannual memory' means or how it was determined if a catchment displayed this.

Original text:

"Catchment memory: we selected catchments that exhibit minimal interannual memory (as defined by de Lavenne et al., 2022), because the simple annual model used here cannot reproduce interannual memory effects."

Interannual memory refers to the persistence of hydrological conditions from one year to the next; for example, when the storage state at the end of year *n* significantly influences streamflow in year *n+1*. For these catchments, the simple annual regression between annual anomalies of streamflow and climatic characteristics is not enough, and additional terms (i.e., precipitation anomalies of the previous years) are needed. We followed the detection method of de Lavenne et al. (2022), which quantifies this memory from observed streamflow autocorrelation. We will add a brief definition at this point in the manuscript.

L215 — PET anomaly and synchronicity as predictors

The PET anomaly (and to a lesser extent the synchronicity anomaly) are unlikely to be significant predictors of annual streamflow. Results in Figure 4 seem to bear out this assumption. I have to wonder if the additional complexity introduced by using these parameters is worthwhile.

Original text:

" Traditionally, the most important factors used to explain streamflow elasticity are (i) the annual precipitation anomaly (ΔP_n) and (ii) the annual potential evaporation anomaly ($\Delta {E_0}_n$). Following Andréassian et al. (2025), we incorporated a third factor to explain streamflow elasticity, namely, the synchronicity between precipitation and potential evaporation Λ_n. This factor represents the volume of annual precipitation that is easily accessible for evaporation."

This is a legitimate question. You are correct that Figure 4 shows that the synchronicity and PET elasticity coefficients are often statistically non-significant and set to zero in many catchments. However, retaining these predictors in the model structure serves two purposes. First, for the catchments where they are significant, they provide additional explanatory power at no cost in terms of model structure (the coefficients are simply zeroed out where non-significant). Second, and more importantly, including these terms follows Andréassian et al. (2025), who showed that synchronicity in particular captures a physically meaningful process — the temporal overlap between precipitation and potential evaporation — that can matter in specific climate contexts. We will add a sentence in the discussion to acknowledge the reviewer's observation and clarify the rationale for retaining these terms.

L257 — "Data are not treated as time series"

It is stated that the data are not treated as time series, but the model is predicting annual Q (not MAQ). Surely then, it is a time series?

Original text:

"It should be noted that this model does not account for temporal dependencies, as the data are not treated as time series (hence the exclusion of catchments with significant memory)."

The apparent contradiction is understandable. The model does produce annual streamflow estimates for successive years, but no temporal autocorrelation structure is assumed or exploited between years; each year is treated as an independent observation during training, with no recurrent or memory-based mechanism linking year *n* to year *n+1*. We will rephrase this sentence and add a diagram explaining the structure of the model to remove the ambiguity.

L261 — Three subsets: spatial or temporal?

Do you mean that the catchments were separated into three subsets (i.e. spatially), or the data for each catchment was separated into three subsets (i.e. temporally)?

Original text:

"The dataset was partitioned into three distinct subsets: a training set to calibrate the model, a validation set to optimize hyperparameters, and a test set to assess the model's performance on unseen data."

Both interpretations apply, depending on the cross-validation experiment considered. In the temporal experiment (HYBRID TEMP), the three subsets correspond to non-overlapping time periods applied to all catchments. In the spatial experiment (HYBRID REGIO), the three subsets correspond to disjoint groups of catchments across all available years. In the spatiotemporal experiment (HYBRID SPATIO-TEMP), both dimensions are split simultaneously. This sentence introduces the general framework; the specific partitioning strategy for each experiment is described in detail in Sect. 3.4. We will clarify this at L261 to avoid confusion.

L315 — "Five groups of similar size"

Does this mean that catchments in each group were of significantly different size (i.e. area)? Please make this clearer.

Original text:

"For spatial cross-validation (regionalization), catchments were randomly divided into five groups of similar size."

"Similar size" refers to the **number of catchments** in each group (~600 each), not to their physical drainage area. The groups were assigned randomly without regard to catchment area. We will rephrase to remove the ambiguity.

We hope these responses fully address the reviewer's comments and that the revised manuscript will be suitable for publication.