Review of the paper entitled : « Sensitivity-Aware Gradient Estimation (SAGE) for Rapid Continental-Scale Training of Hydrologic Models"

Dear Authors,

This manuscript addresses a highly relevant question: how to reduce reliance on standard automatic differentiation (AD) for the full gradient chain in hybrid (hydrological) models. While the broader field of hybrid modeling is rapidly advancing through the widespread adoption of modern AD libraries, this study offers high-quality, methodological research focused on the core problem of gradient computation of a hybrid forward model composed of a descriptor-to-parameters neural network (NN) chained to lumped hydrological models consisting in a set of ordinary differential equations.

The study decomposes the gradient of the loss wrt the trainable parameters (here, the NN weight for regionalized prediction of conceptual hydrological model parameters from descriptors) into 3 general terms via the chain rule: the NN Jacobian ($\partial n/\partial W$), the hydrological Jacobian ($\partial \boldsymbol{q}/\partial \boldsymbol{\theta}$), and the loss-sensitivity term ($\partial \ell/\partial \boldsymbol{q}$). The first two terms are computed by analytical derivation, while the last is computed as described in their previous study (Vrugt et al. 2026 preprint). The core question is how and when to efficiently couple gradients with analytical formulations in hybrid models. A similar question was posed and method proposed in Huynh et al. (2024, 2025), though simpler, by coupling the exact regionalization NN Jacobian with adjoint-based gradients for regionalization of spatially distributed parameters and spatio-temporal flux correction learning. The proposed framework is demonstrated on several well established lumped hydrological models and in descriptor based regionalization on a set of around 500 US catchments from CAMELS dataset.

Overall, while this contribution is highly relevant, solid and rich, the manuscript is currently overly long and several comments should be addressed before publication, especially regarding lit positioning, methodological clarification, and generability/transferability. Details are provided below.

Best regards,

Pierre-André Garambois and Ngo Nghi Truyen Huynh

Major Comments

- Very valuable lightweight sensitivity equation of spatially lumped hydrological ODE. Need to refine the positioning of the proposed method that I think is applicable to semi/spatially distributed models, with independent "runoff produciton" operators on each cells on top of (differentiable or not) routing, to various hybridaiton with NN that do not depend on hydrological model state.
- Sensitivity equations compute forward model output derivatives wrt to the input (see early forward sensitivity approaches in low dimensional param spaces in Guinot and Cappelaere (2009)) with hydraulic hyperbolic PDE) while adjoint methods compute gradients indirectly through an auxiliary (adjoint) system, exploiting the chain rule in reverse.
- The authors spent many places to discuss/criticize about modern AD and finite differences however the view of AD is still not precise somewhere. For example in

line 31-39, Feng et al. and Shen et al. used AD within DL frameworks like PyTorch; they do not use finite-difference approximations in their frameworks. To the best of our knowledge, no modern hybrid hydrological modeling framework trains NNs using finite differences (relative “high dimensional” problem). Another example is the claims in line 58-72 that reverse-mode gradient tracking is "conceptually distant" from hydrologic practices, “difficult to inspect or interpret”, and has “nontrivial overhead in terms of memory usage” seem to overlook the rich history of adjoint-based methods (continuous/discrete reverse-mode) in variational data assimilation (it is possible to analytically derive adjoint models, or obtain the exact adjoint source code using AD). Recent frameworks specifically combine the NN Jacobian with the gradients of hydrological components obtained via adjoint models. Building directly on the longstanding development of gradient-based optimization in geophysical modeling, adjoint-based methods resolve the exact issues the authors raise: they are mathematically transparent (fully inspectable) and completely bypass the massive memory overhead of modern AD libraries using dynamic computational graphs. Instead of presenting a dichotomy between AD and finite differences, it would be far more accurate and valuable to distinguish between modern AD within standard DL frameworks and adjoint-based variational (hybrid) frameworks.

- Mention numerical adjoints of hydrological/hydrodynamic model (e.g. Castaings et al. 2009, Monnier et al. 2016). Furthermore, you could clarify that the cited hybrid methods and also this study are based lumped parameters by catchment, which differs from fully spatially distributed approaches.
- Generality, transferability, and implementation cost: The SAGE framework maps attributes to parameters outside the dynamical loop. An open question is what happens when neural networks are embedded directly within the model structure, for instance when NN weights act as parameters controlling dynamic flux corrections, or when they are integrated within universal differential equations (UDEs) or state-space formulations. In such cases, tracking forward sensitivities for a large number of internal, potentially time-varying NN weights could lead to a significantly enlarged augmented ODE/PDE system, with substantial computational cost. Clarification on the limits of transferability in such configurations would be helpful.
  (Out of curiosity, how would this scale in relation to recent work cited by the authors (Huynh et al., 2026), which relies on analytical gradients to compute Jacobians of state-dependent neural networks (learned closures) within an implicit UDE solver?).
  Furthermore, regarding implementation cost: since the hydrological Jacobian term is analytically derived, modifications to the physical model structure may require re-derivation and re-implementation of the corresponding equations. The authors are encouraged to discuss the practical implications of this.

*****

Minor Comments:

Abstract & Introduction

- **Abstract (Naming):** The acronym **SAGE** might inadvertently cause minor confusion with the well-known software library SageMath.
- **Abstract (Gradient Clarification):** In the sentence: *“Compared to conventional training strategies based on numerical differentiation or automatic differentiation,*

SAGE achieves machine-precision agreement with reference gradients while reducing computational cost by several orders of magnitude..." please explicitly clarify how the adjoint/gradient computation is handled.

- **Line 26:** Please clarify if the cited works here are exclusively pure machine learning (ML) hydrological models.
- **Line 27:** Consider explicitly positioning traditional calibration methods against the high-dimensional inverse problems encountered in pure deep learning (DL) or hybrid models, as well as in spatially distributed models.
- **Line 31 (State of the Art):** Enrich the literature review and positioning by including regionalization of spatially distributed models using learnable neural network (NN) mappings between descriptors and parameters. Specifically, cite the gradient chain rule with model sensitivity from *Huynh (2024, WRR)*, hybrid models with state-dependent NNs from *Huynh (2025, HESS)*, and explicit Jacobians from *Huynh (2026, GMD)*.
- **Line 45 (Forward Sensitivity):** Consider adding relevant references for forward sensitivity from the mathematical/applied mechanics community (e.g., *Delenne et al., 2011* regarding the sensitivity of 1D shallow water equations with source terms) and hydrological finite-difference sensitivity framework works (e.g., *Gupta and Razavi, 2018*).
- **Line 60:** Modify this section to contextualize your work against existing studies that already implement analytical derivations of NN gradients and the chain rule—specifically where gradients are computed via a numerical adjoint for the remaining spatially distributed nonlinear routing components. Also, discuss state-dependent NNs (e.g., *Huynh 2026, GMD*).
- **Line 62:** The phrase *"conceptually distant"* is unclear. It might be more accurate to state that modern libraries like Jax and PyTorch impose strict coding and algorithmic rules to ensure computationally efficient adjoint generation (similar to how frameworks like Tapenade require specific coding structures to adjointize Fortran code).
- **Line 70:** The statement *"AD provides exact derivatives"* should be refined. Automatic differentiation provides **machine-precision** derivatives rather than purely analytical ones.

Section 2 (Methodology & Equations)

- **Line 100:** Please include explicit citations for the works mentioned here.
- **Line 104:** Define clearly. It would improve readability to place the model definition and its dependencies in a standalone display equation rather than inline.
- **Line 105:** Clarify the term *"pointwise cost"* and polish the surrounding sentence for clarity.
- **After Line 106 (The Loss Function):** The text states that the formulation works for both pointwise loss functions (e.g., MAE, MSE) and reward-based goodness-of-fit metrics (NSE, KGE). However, the provided formulation is only valid for separable, pointwise loss functions. Because metrics like NSE and KGE rely on global statistics (such as variance across the entire time series), they are non-separable. Please reformulate this equation to be more generic. Additionally, explicitly note that the chosen cost function must be differentiable; certain widely used evaluation metrics (e.g., KGE) contain non-differentiable points.
- **Equation 1:** This represents an affine scaling function. Please explicitly define the new bounds following transformation—specifically, the scaling from the

normalized space to the physical parameter bounds . Note that integrating a bijective, differentiable mapping with a bounded image (e.g., a sigmoid function) within the forward model was previously proposed by *Huynh et al. (2024)*.

- **Line 120:** Please clarify exactly what is meant by *"state equations augmented with sensitivity equations"* (e.g., briefly explain how the state vector and sensitivity matrix are concatenated for the ODE solver). Presenting this step via a standalone display equation involving the numerical solver would be highly beneficial.
- **Notation Check (Line 120):** Ensure that the notation is consistent. Is for simulated discharge from the numerical solver using the same notation as observed discharge introduced at the beginning of the section? Conversely, check if is being mixed up. Please clarify and simplify.
- **Link between Eq. 2 and Eq. 3:** The transition here feels disjointed. Equation 2 defines the scaling Jacobian, but Equation 3 jumps directly into the loss gradient with respect to physical parameters. Mathematically, they are not connected in the text until Equation 5. I recommend restructuring this sequence to improve logical flow.
- **Line 125 / Eq. 3:** Why does the loss-sensitivity term depend notationally on if Line 128 states that it *"depends exclusively on the form of the loss function"*? While the numerical values clearly rely on the simulated discharge (which itself depends on), the current phrasing is slightly contradictory. Please clarify how discrepancies between observations and simulations are propagated back through the model.
- **Line 138 / Eq. 4:** The use of the "conus" subscript is quite specific. Consider using a more generic subscript name for broader applicability.
- **Line 155:** The phrase *"which are then rescaled to physical"* should be adjusted. It is more accurate to say **mapped** rather than *rescaled* when describing a descriptor-to-parameter artificial neural network (ANN) function.
- **Equation 5:** Please explicitly recall what each term represents, and clarify what parameter or variable the phrase *"by repeated application of chain rule"* is operating with respect to.
- **Dimensionality (Eq. 5):** Providing the exact dimensions of the Jacobian matrices in the text immediately below the equation would greatly assist the reader.
- **Line 160:** The definition is redundant as it was already defined above.
- **Line 163:** The phrase *"without finite differences or automatic differentiation of the hydrologic core"* should be clarified to state that these are exact analytical calculations paired with the numerical integration of an ODE.
- **Section 2.2:** Deriving the analytical Jacobian of a simple feed-forward neural network is a classical, well-known operation. To improve the manuscript's flow, this section should be simplified or moved to an appendix. For visual lightness, standardizing the notation to use for weights and biases is recommended.
- **Equation 8 and Line 180:** Define the network Jacobian by catchment along with its dimensions. Reviewing this notation to use might be more precise and contextually relevant.
- **Section 2.5:** Please briefly remind the reader what represents in this context.

Sections 3–5 & Conclusions

- **Line 340:** This explanation should be expanded or summarized more clearly. While using an Empirical Cumulative Distribution Function (ECDF) allows the reader to visualize the full distribution of basin performances, the phrase *"facilitate the objective ranking based on their entire NSE distribution"* seems too strong given that the scalar score ultimately summarizes the distribution via its first moment.

- **Section 5 (Coding Details):** The implementation and coding details in Section 5 are technical and would be better suited for an appendix to keep the primary narrative focused.
- **Line 531:** *"Nonetheless, the overall computational demands remain modest compared to the burden typically associated with automatic differentiation-based training frameworks..."* Please provide at least an approximate order of magnitude for these savings to ground this claim.
- **Line 745:** Rather than the broad claim that *"SAGE enables fast, stable, and transparent large-sample learning,"*it would be more accurate to specify that it enables descriptor-to-parameter regionalization learning for lumped models.
- **Line 758:** *"In practical terms, SAGE reduces large-sample training from days to minutes for CONUS-scale experiments."* Please provide citations or concrete benchmarking data from your study to substantiate this timeframe.
- **Line 838:** Provide a brief, accessible explanation of the LASSO method for readers who may not be deeply familiar with regularized regression.
- **Line 851:** Replace or simplify the word *"idiosyncrasies"* to ensure the text is easily understood.

Figures & General Presentation

- **Figure 1:** This diagram is currently overloaded with detailed equations, which obscures the overarching logic connecting the four primary blocks. Additionally, please clarify what the different color codes signify.
- **Figures 4 and 10 (GR4J Performance):** The authors need to address the abnormal behavior of the cost values for the GR4J model. The cost curves do not reflect a well-trained, converging model. Please discuss whether this is caused by numerical instability, initialization issues, or suboptimal hyperparameters (such as an improper learning rate).
- **Figure 4 & 10 Clarification:** For the validation scores plotted across iterations, please clarify in the title of subplot (b) if these are calculated using the trained parameters to avoid reader confusion.
- **Figure 5:** The poor performance of the GR4J model is surprising, as it is traditionally a highly robust performer across the CONUS dataset. Please provide context or an explanation for this deviation.
- **Figure 8:** Please add context or comparison regarding how this approach positions itself relative to existing VIC-related (Variable Infiltration Capacity) regionalization studies.
- **Parameter Maps (General Interpretation):** Why are the parameter maps displayed using normalized values? For physical interpretability and to properly assess the orders of magnitude of the learned parameters, it would be far more valuable to display the actual physical values.

References

Castaings, W., Dartus, D., Le Dimet, F.-X., & Saulnier, G.-M. (2009). Sensitivity analysis and parameter estimation for distributed hydrological modeling: potential of variational methods. Hydrology and Earth System Sciences, 13, 503–517. https://doi.org/10.5194/hess-13-503-2009

Guinot, V., & Cappelaere, B. (2009). Sensitivity analysis of 2D steady-state shallow water flow: Application to free surface flow model calibration. Advances in Water Resources, 32(4), 540–560. https://doi.org/10.1016/j.advwatres.2009.01.005

Huynh, N. N. T., Garambois, P.-A., Colleoni, F., and Monnier, J. (2026) A hybrid physics–AI approach using universal differential equations with state-dependent neural networks for learnable, regionalizable, spatially distributed hydrological modeling, Geosci. Model Dev., 19, 1055–1074, https://doi.org/10.5194/gmd-19-1055-2026

Huynh, N. N. T., Garambois P.-A, Renard, B., Colleoni F., Monnier J., Roux H. (2025) A distributed hybrid physics–AI framework for learning corrections of internal hydrological fluxes and enhancing high-resolution regionalized flood modeling. Hydrol. Earth Syst. Sci. https://doi.org/10.5194/hess-29-3589-2025

Huynh, N. N. T., Garambois P.-A, Renard, B., Roux, H., Demargne J., Javelle P. (2024) Learning Regionalization Using Accurate Spatial Cost Gradients Within a Differentiable High-Resolution Hydrological Model: Application to the French Mediterranean Region. Water Resources Research. https://doi.org/10.1029/2024WR037544

Lee, H., Seo, D.-J., Liu, Y., Koren, V., McKee, P., & Corby, R. (2012). Variational assimilation of streamflow into operational distributed hydrologic models: effect of spatiotemporal scale of adjustment. Hydrology and Earth System Sciences, 16, 2233–2251. https://doi.org/10.5194/hess-16-2233-2012

Monnier, J., Couderc, F., Dartus, D., Larnier, K., Madec, R., & Vila, J.-P. (2016). Inverse algorithms for 2D shallow water equations in presence of wet–dry fronts: Application to flood plain dynamics. Advances in Water Resources, 97, 11–24. https://doi.org/10.1016/j.advwatres.2016.07.005