Dr. Nunzio Romano
Editor, *Hydrology and Earth System Science*
08-17-2026

Dear Dr. Romano (Nunzio):

Please find attached our response to the first-round review of our manuscript entitled "Sensitivity-Aware Gradient Estimation (SAGE) for Rapid Continental-Scale Training of Hydrologic Models" (HESS Tracking No. #egusphere-2026-693). We sincerely appreciate the thoughtful and constructive comments provided by the two reviewers. Their feedback has been very helpful in identifying aspects of the manuscript that require clarification, and additional context/development.

We have carefully considered all comments and provide point-by-point responses below. If the editor invites a revised manuscript, we would focus on the following main areas:

(i) clarify the motivation for differentiable hydrology, including the roles and differences of finite-difference, automatic-differentiation, and sensitivity-based approaches;
(ii) improve the accessibility and organization of the manuscript by streamlining theoretical developments, clarifying notation and terminology, strengthening transitions between sections, and providing additional introductory guidance for the case studies;
(iii) expand discussion of forward sensitivity analysis, including its strengths, limitations, trade-offs, and relationship to automatic differentiation and large-sample applications; and
(iv) clarify the case-study data, experimental design, optimization behavior, and interpretation, while removing comparisons not supported by the submitted experiments.

The accompanying document responds to each comment and distinguishes clarifications that can be made in a possible revision from requests that would require separate research beyond the scope of this paper. No revised manuscript has been prepared at this final-response stage. If revision is invited, we will address the corrections and focused clarifications identified below and will provide a tracked-changes version.

The software associated with this work will be made publicly available at https://github.com/jaspervrugt/diffhydrology upon formal acceptance of the paper. In addition, a companion software paper describing the standalone SAGE-GUI graphical user interface is currently under review [3].

We thank the two reviewers again for the time and care they have invested in evaluating our manuscript. Their comments provide valuable guidance for a possible revision while also identifying several broader questions that are best treated as future research.

Sincerely,

Jasper A. Vrugt

COMMENTS FROM REVIEWERS

The following pages contain our point-by-point responses to the reviewers' comments. Where appropriate, we identify focused changes that we would make if invited to revise; no revised manuscript has been prepared at this stage.

Reviewer #1: Pages 3-15
Reviewer #2: Pages 16-19
Our responses are given in blue and regular font following each of the reviewers' comments.

Reviewer #1

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 ($\partial \mathbf{q}/\partial \boldsymbol{\theta}$), and the loss-sensitivity term ($\partial \mathcal{L}/\partial \mathbf{q}$). 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

RESPONSE: We thank the reviewers for their careful reading and constructive assessment. The comments help distinguish focused clarifications suitable for a possible revision from extensions that would require dedicated research. Our responses below follow that distinction and do not presume that a revised manuscript already exists.

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.

RESPONSE: We agree that semi-distributed and spatially distributed formulations are important. They are not, however, demonstrated in this paper, which studies catchment-lumped ODE models with a static descriptor-to-parameter mapping outside the dynamical loop. Extending SAGE to cell-wise production operators and routing would require dedicated derivation, implementation, and benchmarking to establish when the augmented sensitivities remain tractable and how routing contributes to the gradient. We therefore regard this as out of scope rather than promise such a revision. If invited to revise, we will make the demonstrated lumped-model scope explicit and identify distributed extensions as future research.

- 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.

RESPONSE: We agree with this general distinction: forward sensitivity analysis propagates state and output derivatives with respect to parameters alongside the forward solution, whereas an adjoint evaluates the required gradient through an auxiliary reverse system. The comment does not identify an error in a particular equation, statement, or figure, so no specific manuscript change is warranted on this point alone. If revision is invited, we will nevertheless check that this distinction is stated accurately where the two approaches are compared.

- 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 transpar-

ent (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.

RESPONSE: We appreciate this correction. The submitted introduction groups distinct approaches too broadly and overstates the opacity and overhead of reverse-mode methods. PyTorch-based hybrid studies use AD, not finite differences, and continuous or discrete adjoints have a long and mathematically transparent history in variational geophysical modeling. If invited to revise, we will distinguish finite differences, AD in modern learning libraries, and continuous/discrete adjoints; remove categorical claims that adjoints are intrinsically opaque; and state that memory, transparency, and implementation costs depend on the realization. This is a warranted correction of the paper's positioning, not a claim that SAGE generally dominates adjoint methods.

- 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.

RESPONSE: The distinction between catchment-lumped and fully spatially distributed parameterizations is relevant and, if revision is invited, we will state it more explicitly. We will also consider the cited numerical-adjoint papers where they directly support the historical or methodological discussion. Citations will be selected for relevance to specific statements rather than added solely because they were suggested.

- 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 rederivation and reimplementation of the corresponding equations. The authors are encouraged to discuss the practical implications of this.

RESPONSE: This is an important limitation. The demonstrated formulation is best suited to a

moderate-dimensional hydrologic parameter vector produced outside the dynamical loop. A state-dependent neural network embedded in the dynamics would require derivatives of the learned component and its coupling to model states; propagating sensitivities with respect to many internal weights could enlarge the augmented system enough that AD or adjoint methods are preferable. Likewise, changing the physical model structure requires rederiving, implementing, and testing the affected sensitivities. A scaling comparison with the Huynh et al. (2026) formulation would require a matched implementation and is new research. If invited to revise, we will add this scope and implementation-cost discussion, but we will not claim or implement the proposed UDE extension in this paper. It may be worth noting that our methodology was designed to closely follow the, so called, differentiable modeling methodology popularized by Shen et al., but we certainly acknowledge that there are many methods of physics-informed ML worth investigating. We will make in effort in the revision to give a more broad context of physics-informed ML and describe more carefully the context of our contribution.

*****

Minor Comments:

Abstract & Introduction

• Abstract (Naming): The acronym SAGE might inadvertently cause minor confusion with the well-known software library SageMath.

RESPONSE: We acknowledge the possible association with SageMath, but do not consider it likely to cause substantive confusion because SAGE is expanded as "Sensitivity-Aware Gradient Estimation" at first use and the subject matter is distinct. We therefore propose to retain the acronym; no manuscript change is needed.

• 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.

RESPONSE: The requested clarification is warranted. If invited to revise, we will state briefly in the abstract that the loss derivative, hydrologic parameter sensitivities, parameter mapping, and network Jacobian are combined through the chain rule. We will also remove or qualify comparisons with numerical differentiation or AD that are not supported by a matched benchmark in this paper.

• Line 26: Please clarify if the cited works here are exclusively pure machine learning (ML) hydrological models.

RESPONSE: We agree that the current grouping can be read too broadly. If invited to revise, we

will identify which cited examples are purely data driven and which are hybrid, without expanding the discussion beyond what is needed for accurate positioning.

• 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.

RESPONSE: This is a useful but small framing clarification. If invited to revise, we will distinguish conventional low-dimensional basin calibration from the higher-dimensional inverse problems created by neural-network training and some distributed parameterizations.

• 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 statedependent NNs from Huynh (2025, HESS), and explicit Jacobians from Huynh (2026, GMD).

RESPONSE: The submitted paper should acknowledge directly relevant prior gradient constructions, including work on learnable regionalization and state-dependent networks, where this is needed to delimit novelty. If invited to revise, we will cite the relevant Huynh et al. papers only in support of specific comparisons and explain the architectural difference from SAGE. We will not characterize those methods as simpler without a precise basis, nor add citations solely because they were requested.

• 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).

RESPONSE: We appreciate these references. This is the second paper in a series, where the first paper was more focused on the theoretical background [4] with more historical and complete references of FSA, and this present paper is demonstrating the methodology applied to regional/large-scale training. If invited to revise, we will consider adding those that directly support the historical discussion of forward sensitivity analysis. Gupta and Razavi's finite-difference framework is methodologically distinct from analytic forward sensitivity equations, so any citation would preserve that distinction.

• 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).

RESPONSE: We agree that existing combinations of explicit neural-network derivatives and adjoint-derived physical-model gradients are relevant to positioning. If invited to revise, we will acknowledge them concisely and clarify that this paper treats a static descriptor-to-parameter mapping external to a lumped dynamical model. A full review or implementation of state-dependent distributed architectures is outside scope.

• 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).

RESPONSE: We agree that "conceptually distant" is vague and unnecessarily dismissive. If invited to revise, we will remove it and describe the concrete implementation constraints, as you accurately describe, with appropriate qualification.

• Line 70: The statement "AD provides exact derivatives" should be refined. Automatic differentiation provides machine-precision derivatives rather than purely analytical ones.

RESPONSE: We agree that "exact" requires qualification. If invited to revise, we will use "derivatives of the implemented numerical program evaluated to floating-point precision" or equivalent wording, and apply the same care to claims about SAGE sensitivities, which are also affected by numerical integration tolerances and floating-point arithmetic.

Section 2 (Methodology & Equations)

• Line 100: Please include explicit citations for the works mentioned here.

RESPONSE: If revision is invited, we will add citations for FSA including [1].

• Line 104: Define clearly. It would improve readability to place the model definition and its dependencies in a standalone display equation rather than inline.

RESPONSE: This is a presentation suggestion rather than a scientific correction. If invited to revise, we will improve the definition and will use a display equation if it materially improves readability.

• Line 105: Clarify the term "pointwise cost" and polish the surrounding sentence for clarity.

RESPONSE: We agree that "pointwise cost" should be defined or avoided. If invited to revise, we will clarify that it denotes a loss contribution evaluated at an individual time step and polish the sentence.

• 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.

RESPONSE: We agree; this is a substantive mathematical clarification. If invited to revise, we will write the objective generically as a differentiable function of the complete simulated and observed series, distinguish separable losses from non-separable metrics such as NSE and KGE, and state that the required derivatives must exist at the evaluated point. We will not imply that a sum of pointwise terms represents every metric.

• 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).

RESPONSE: We agree that Equation 1 should identify the normalized and physical bounds explicitly. If invited to revise, we will call it an affine mapping and define both intervals. Other differentiable bounded mappings are possible, but the manuscript need not survey alternatives; we will cite prior use only if it directly informs this statement.

• 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.

RESPONSE: This clarification would improve accessibility. If invited to revise, we will state that the physical states and parameter-sensitivity matrix form an augmented state integrated simultaneously by the numerical solver, including dimensions or a compact equation where helpful.

• 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.

RESPONSE: We will check the submitted notation carefully. If revision is invited, any instance that conflates observed and simulated discharge will be corrected and the two symbols will be defined once and used consistently.

• 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.

RESPONSE: We agree that the transition can be clearer. If invited to revise, we will connect the mapping Jacobian explicitly to the later chain-rule product; we will reorder equations only if needed to improve the 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.

RESPONSE: The reviewer is correct. The functional form of the loss derivative is determined by the selected objective, whereas its evaluated value depends on simulated and observed discharge and therefore indirectly on the parameters. If invited to revise, we will correct this wording and explain that the resulting loss derivative is propagated through the hydrologic-model Jacobian.

• Line 138 / Eq. 4: The use of the "conus" subscript is quite specific. Consider using a more generic subscript name for broader applicability.

RESPONSE: This is a reasonable notation improvement. If invited to revise, we will replace the application-specific subscript in the general derivation with a generic ensemble notation.

• 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.

RESPONSE: We agree. If invited to revise, we will use "mapped" in this context.

• 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.

RESPONSE: If invited to revise, we will define each factor adjacent to Equation 5 and state explicitly that the repeated chain rule yields the loss derivative with respect to trainable network weights.

• Dimensionality (Eq. 5): Providing the exact dimensions of the Jacobian matrices in the text immediately below the equation would greatly assist the reader.

RESPONSE: We agree that dimensions would make the factorization easier to verify. If invited to revise, we will provide the dimensions compactly below Equation 5 and check their consistency.

• Line 160: The definition is redundant as it was already defined above.

RESPONSE: On line 122 a general description of Jacobian for the ODE system is made, while on line 160 the Jacobian for the ANN defined. We do agree that there is some redundancy here, and in other places in the initial submission. We will remove the repetition in a possible revision.

• 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.

RESPONSE: We agree. The sensitivity right-hand sides are derived analytically but integrated numerically with the model ODEs; the evaluated sensitivities therefore remain subject to solver tolerances and floating-point arithmetic. If invited to revise, we will state this explicitly and avoid unqualified use of "exact."

• 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.

RESPONSE: We agree that the feed-forward-network Jacobian is standard. We included it to make the complete factorization self-contained, but it need not dominate the presentation. If invited to revise, we will shorten the derivation or move algebraic detail to an appendix while retaining the definitions needed to reproduce Equation 5; we will also standardize the notation for weights and biases.

• 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.

RESPONSE: If invited to revise, we will define the network Jacobian for an individual basin, give its dimensions, and use the basin index consistently.

• Section 2.5: Please briefly remind the reader what represents in this context.

RESPONSE: The symbol is missing from the extracted reviewer text, but we will check the referenced passage. If revision is invited and the symbol is not locally clear, we will add a brief reminder.

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.

RESPONSE: We agree that the ECDF displays the full cross-basin distribution whereas the scalar score summarizes it. If invited to revise, we will remove the stronger claim of ranking based on the entire distribution and describe precisely what information the scalar retains.

• 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.

RESPONSE: We appreciate the presentation suggestion. The implementation details support reproducibility, but some may interrupt the main narrative. If invited to revise, we will streamline Section 5 and move only lower-level material that is not needed to interpret the experiments to an appendix.

• 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.

RESPONSE: We agree that the submitted experiments do not support a quantitative comparison with AD because no matched implementation was tested. Conducting such a benchmark would be a substantial additional experiment and is outside the present study. If invited to revise, we will remove the implied comparison and report only measured runtime and memory for SAGE, clearly separated from any hypothesis about other implementations.

• 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.

RESPONSE: We agree. If invited to revise, we will narrow this claim to descriptor-to-parameter regionalization for the catchment-lumped models and experiments actually demonstrated.

• 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.

RESPONSE: The "days to minutes" comparison is not supported by a matched benchmark in this manuscript. If invited to revise, we will remove it and report only the observed runtime of the stated SAGE experiments.

• Line 838: Provide a brief, accessible explanation of the LASSO method for readers who may not be deeply familiar with regularized regression.

RESPONSE: If invited to revise, we will add a brief explanation that LASSO uses an ℓ_1 penalty, which can shrink weak coefficients to zero while fitting the regression.

• Line 851: Replace or simplify the word "idiosyncrasies" to ensure the text is easily understood.

RESPONSE: We agree. If invited to revise, we will replace "idiosyncrasies" with a more direct description of the model-specific behavior.

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.

RESPONSE: We prefer to retain Figure 1 because the equations connect the four blocks and make the gradient construction self-contained. We agree that the color coding should be explicit. If invited to revise, we will improve the caption and consider modest visual simplification without removing information essential to the workflow.

• 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).

RESPONSE: We agree that the submitted GR4J traces do not show satisfactory convergence and should not be explained post hoc by an untested single cause. In the past months we have implemented a more robust version of the Adam/AdamW algorithm and this implementation with gradient clipping and learning rate annealing results in a much smoother decay of the loss function. If invited to revise, we will update the loss function training/evaluation plots. Unexpected (oscillatory/non-convergent) behavior will be discussed and inspected in relationship to initialization, learning rate, parameter bounds, numerical behavior, and optimization settings.

• 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.

RESPONSE: This is a useful caption clarification. If invited to revise, we will state whether each validation score uses the parameter mapping associated with the network weights at that iteration.

• 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.

RESPONSE: We agree that the result requires context. Figure 5 evaluates this continuous-time, extended GR4J implementation under a shared static descriptor-to-parameter mapping and the stated optimization setup; it is not evidence of the best performance attainable by standard individually calibrated GR4J. The submitted traces also show problematic optimization. If invited to revise, we will make these distinctions explicit and avoid assigning the performance gap to one cause without a controlled diagnosis.

• Figure 8: Please add context or comparison regarding how this approach positions itself relative to existing VIC-related (Variable Infiltration Capacity) regionalization studies.

RESPONSE: The regionalization benefits from differentiable modeling has been described at length by [2]. We are not claiming that we have a novel method of regionalization, rather, we are using Figure 8 to demonstrate the successful application of analytic gradients to regionalization.

• 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.

RESPONSE: The normalized maps are appropriate for comparing relative spatial structure across parameters with different units and ranges, but they do not by themselves support interpretation of physical magnitudes. Producing a full additional set of physical-unit maps is not necessary to evaluate the gradient method. If invited to revise, we will explain why normalized values are shown and provide the physical mapping bounds or a compact physical-unit summary so readers can recover magnitudes. We will consider physical-unit panels only where they materially aid interpretation.

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/hess13-503-2009

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

Reviewer #2

This manuscript presents SAGE, a framework that combines neural-network-based parameter regionalization with analytically derived model sensitivities for large-sample hydrologic model training. The method is demonstrated using six conceptual rainfall–runoff models across 531 CAMELS basins.

RESPONSE: We thank the reviewer for the concise summary and careful evaluation of the manuscript.

Overall, I believe the method could make a useful contribution. I appreciate the practical idea of retaining an established C++ hydrologic model while still enabling end-to-end gradient-based training. This provides a pathway for future development of mature, non-Python hydrologic models, including modifying or replacing selected process components for hypothesis testing and scientific discovery. My main concern is that some of the broader claims regarding efficiency, predictive accuracy, and interpretability are not yet fully demonstrated. I have the following comments.

RESPONSE: We agree that several claims in the submitted manuscript extend beyond the evidence shown. If invited to revise, we will narrow claims about efficiency, predictive accuracy, and interpretability and distinguish measured results from prospective capabilities. Our earlier paper provides complementary derivative comparisons, but it does not substitute for a matched AD benchmark of the present experiment.

Major comments

1. The efficiency claim needs to be supported by a matched benchmark. The manuscript discusses the runtime and memory limitations of reverse-mode automatic differentiation (AD). However, I did not see a matched comparison with an equivalent PyTorch or JAX implementation. Some of the efficiency discussion relies on general statements such as training taking "days," but no direct empirical evidence is provided under the same experimental setup. Ideally, the two approaches would use the same model equations, ANN architecture, basin set, training period, loss function, hardware, and convergence criterion. A comparison using one representative hydrologic model would already be informative. I suggest reporting wall-clock runtime, peak memory use, convergence behavior, and final model performance for both approaches. This would provide a clearer basis for evaluating the actual efficiency gain offered by SAGE.

RESPONSE: We agree that a matched implementation is required to quantify a speed or memory advantage over PyTorch or JAX. Building and validating an equivalent implementation under identical equations, data, hardware, stopping criteria, and metrics would be a substantial study in its own right and is outside the scope of this paper. If invited to revise, we will remove the "days"

and orders-of-magnitude comparisons, report only the measured wall-clock and memory behavior of SAGE, and identify a matched benchmark as future work rather than promise one here.

2. The practical scope of SAGE should be clarified. One potential advantage of SAGE is that it can work with hydrologic models implemented in compiled languages without requiring the full model to be rewritten in PyTorch or JAX. However, this advantage comes with the need to derive and maintain the analytical sensitivities. It would therefore be helpful for the authors to explain this trade-off more clearly and when they consider SAGE the preferable choice to guide users. For more complex or spatially distributed models, is the sensitivity derivation still manageable?

RESPONSE: We agree with this trade-off. SAGE is most attractive for a mature compiled model with a moderate number of parameters and sensitivities that can be derived, verified, and maintained. The up-front and maintenance costs increase with structural complexity and spatial dimensionality; models with many distributed parameters or frequent structural changes may favor AD or adjoints. If invited to revise, we will add this practical guidance. Extending and testing SAGE on a distributed model is outside the present paper.

The manuscript also mentions future hybrid physics-aware formulations (line 1044), which is a direction I find appealing for scientific discovery. Could the authors clarify how a neural network component would be incorporated into the SAGE framework? Does it require additional manual derivations? A clearer discussion of these practical limits would help readers understand the types of models and applications for which SAGE is most suitable.

RESPONSE: A network that maps descriptors to a modest parameter vector outside the dynamical loop fits the demonstrated factorization. A state-dependent network embedded in a flux or state equation would require additional derivatives of the network and its coupling to physical states, and forward propagation with respect to many weights could become impractical. This architecture is not demonstrated here. If invited to revise, we will state this boundary and temper the prospective hybrid-language; implementing such a hybrid model is new research and outside scope.

3. The physical model benchmarking performance needs more clarification. The individually calibrated model results in Part A are obtained using L-BFGS-B. Can the SAGE analytical gradients also be used for individual-basin calibration? If so, it would be useful to include a Part A benchmark using SAGE, or at least clarify whether the results in Figure 5 were obtained with SAGE-derived gradients. This would help demonstrate that SAGE itself can reproduce strong physical-model benchmarks across CONUS. Otherwise, applying SAGE only in the regional static-parameter setting makes it difficult to tell whether SAGE's under performance relative to Part A comes from the regional/static parameterization or from a difference in optimization procedure.

RESPONSE: Yes. The hydrologic sensitivities can supply gradients for individual-basin calibration by omitting the descriptor-to-parameter network factor. However, Part A uses the stated L-BFGS-

B procedure and is not a separate SAGE-gradient calibration experiment. Adding that experiment would change the study design and is not needed to describe what Part A establishes. If invited to revise, we will clarify this distinction and state that the Part A–Part B gap cannot be attributed uniquely to regionalization, static parameterization, or optimizer behavior from the reported results.

In addition, I do not think the current benchmarking accuracy supports the statement that the SAGE-trained conceptual models are comparable to the data-driven LSTM benchmark, given that the reported best median performance is approximately 0.66 compared with 0.73–0.75 for the LSTM (Lines 780-785). While the authors later acknowledge that "the SAGE-trained conceptual models remain, on average, below the best machine-learning benchmarks reported for rainfall-runoff simulation.", I suggest softening the earlier statement and describing the performance difference more directly.

RESPONSE: We agree. A median near 0.66 should not be characterized as comparable to reported LSTM medians near 0.73–0.75 without a defined criterion and matched setup. If invited to revise, we will state the numerical difference directly and soften the claim.

On a related note to both accuracy and point 2: if static attributes alone are insufficient to reach high accuracy, does SAGE support a more expressive parameterization network? For example, an RNN taking both dynamic forcing and static attributes, the way dPL uses an LSTM ahead of the physical model skeleton? Is SAGE able to accommodate more advanced architectures and additional inputs for parameter estimation, or is it inherently limited to static-attribute mappings?

RESPONSE: SAGE is not mathematically restricted to the specific feed-forward mapping used here: another differentiable external parameterization can replace it if its Jacobian is available. An RNN driven by dynamic forcing is more consequential because it can introduce time-varying parameters or corrections and additional temporal dependencies inside the dynamical loop. That case is outside the implementation and experiments presented. If invited to revise, we will state this boundary, but we will not claim support for an architecture that has not been derived and tested.

4. The interpretability enabled by the explicit Jacobians should be demonstrated more directly. The manuscript relates the interpretability of SAGE to the availability of explicit Jacobians. However, the current analysis mainly presents spatial maps of normalized parameter values (Figure 8), which can also be produced by other parameter-learning approaches independent of SAGE's sensitivity computation. I suggest including an example that uses the explicit Jacobians for model diagnosis. This would provide a clearer demonstration of the practical interpretability gained from the SAGE formulation.

RESPONSE: We agree that parameter maps do not demonstrate interpretability arising from explicit Jacobians; such maps can be produced by other learning approaches. A dedicated Jacobian-based attribution experiment would be new analysis and is outside the current study. If invited

to revise, we will narrow the claim: explicit Jacobians make local sensitivity pathways available for inspection and diagnosis, but the submitted results do not demonstrate superior physical interpretability. We will identify a basin-level sensitivity analysis as future work rather than imply it has been performed.

Minor comments

Lines 367-369: The calibration and validation periods appear to overlap in water year 1999 (training on WY 1999-2008 while validation over WY 1989-1999). Please clarify the exact years used for spin-up, training, and validation. Do the authors mean WY 2000-2008 for training, and WY 1990-1999 for validation? The statement that "calibrated parameters are often effective quantities" appears twice (first near line 912, again near line 1030). I'd suggest giving the full explanation only where it first appears and removing the repeated version.

RESPONSE: The reviewer has identified a genuine inconsistency: the submitted text labels WY 1999–2008 as a 9-year training period and WY 1989–1999 as a 10-year validation period, although the written endpoints overlap. We will verify the actual experiment settings before changing the dates rather than infer them from the intended durations. If invited to revise, we will correct the spin-up, training, and validation periods consistently in the text, code box, captions, and tables, and remove the repeated explanation of effective parameters.

Line 1040: "sandwich-adjusted uncertainty estimate" would benefit from brief context, or a plain-language note on why this approach delivers "reliable parameter and predictive uncertainties within a single Markov chain Monte Carlo run."

RESPONSE: We agree that "reliable" is too strong. A sandwich adjustment can improve robustness to some forms of misspecification by combining local curvature with variability in score contributions, but a single Markov chain does not by itself guarantee valid parameter or predictive uncertainty. If invited to revise, we will add a short plain-language explanation and qualify the statement with the requirements of adequate sampling, assumptions, and an appropriate adjustment.

References

[1] D. G. Cacuci. Sensitivity theory for nonlinear systems. *Journal of Mathematical Physics*, 22 (12):2794–2802, 1981. doi: 10.1063/1.524385.

[2] C. Shen, A. P. Appling, P. Gentine, T. Bandai, H. Gupta, A. Tartakovsky, M. Baity-Jesi, F. Fenicia, D. Kifer, L. Li, et al. Differentiable modelling to unify machine learning and physical models for geosciences. *Nature Reviews Earth & Environment*, 4(8):552–567, 2023.

[3] J. A. Vrugt and J. M. Frame. SAGE-GUI: A graphical environment for differentiable hydrologic modeling with analytic gradients. *Environmental Modeling & Software*, 2026:XX, 2026.

[4] J. A. Vrugt, J. M. Frame, and E. Bollman. Reclaiming first principles: An analytic differentiable framework for conceptual hydrologic models. *Water Resources Research*, 2026. URL https://arxiv.org/abs/2602.06429. Manuscript under review.