RESPONSE TO REVIEWER #1 FOR HYDROLOGY AND EARTH SYSTEM SCIENCES: MANUSCRIPT EGUSPHERE-2025-6212

BY Zezhen Wu, Zhongwang Wei, Xingjie Lu, Nan Wei, Lu Li, Shupeng Zhang, Hua Yuan, Shaofeng Liu, and Yongjiu Dai

We thank Reviewer #1 for thoughtful and constructive feedback. This Response to the Reviewer file provides a complete documentation of the changes that have been made in response to each individual comment. Reviewer's comments are shown in plain text. Authors' responses are shown in purple color. Quotations from the revised manuscript are shown in blue color.

1. This study introduces the Model Fidelity Metric (MFM) as an alternative to traditional metrics like NSE and KGE. The method demonstrates some practical improvements in specific failure modes, such as error compensation and low-variability conditions, using synthetic tests and the CAMELS dataset. However, the paper requires further improvement in the conceptual explanations, methodological descriptions, and understanding of error metrics. See my comments below.

➔ Thank you very much for your summary. We will address each of your comments and propose revisions to improve our manuscript.

2. Although the study conducted a sensitivity analysis on the hyperparameters, it does not provide specific guidance on their selection. It is suggested to supplement the paper with recommended parameter values or adaptive selection methods to enhance the practical utility of the approach.

➔ Thank you for this important suggestion. We have added a concise practical guide at the end of Sect. 4.5, and a more detailed discussion in Sect. 5.

In Sect. 4.5, we added the following information (P22L501–P22L505):

"We recommend using $p = 1.0$, $n_{\mathrm{SUSE}} = 10$, $n_{\mathrm{PHI}} = 10$, $c = 4.0$ as default settings for general evaluation. When the assessment of extreme events is critical, increasing p to 2.0 imposes heavier penalties on large errors. When finer distributional resolution is needed, increasing n_{SUSE} and n_{PHI} to 100 enables

percentile-level evaluation of variability and distribution similarity. Rather than introducing additional weighting parameters, adjusting these built-in hyperparameters preserves the metric's structural integrity while smoothly modulating sensitivity."

In the Discussion (Sect. 5), we added the following information (P25L560–P25L566):

"We recommend initializing MFM with default parameters ($p = 1.0$, $n_{\mathrm{SUSE}} = 10$, $n_{\mathrm{PHI}} = 10$, $c = 4.0$), where NMAEp is equivalent to NMAE, coarse binning mitigates sparsity, and phase shifts incur mild penalties. For rigorous assessments of extreme events or error modes, or during multi-model comparisons where competitive models achieve close scores and are difficult to distinguish, an enhanced configuration ($p = 2.0$, $n_{\mathrm{SUSE}} = 100$, $n_{\mathrm{PHI}} = 100$, $c = 2.0$) is recommended. This setting equates NMAEp to NRMSE to strictly penalize large errors, employs percentile-based binning for distributional fidelity, and forces ω to 0 under anti-phase conditions via PPF."

3. Errors in land surface variable estimation are usually complex. In many studies, multiple metrics—such as correlation coefficients and bias—are commonly used to better understand the sources of these errors. Individually, these metrics cannot comprehensively reflect model deficiencies, but they offer greater flexibility. For example, soil moisture evaluations tend to emphasize correlation and ubRMSE, with less attention paid to bias. For variables such as ET and LAI, strong seasonality often necessitates decomposing the time series into anomalies and seasonal components, which are then evaluated separately. When developing new error metrics, how do you take these conventional practices into account?

➔ Thank you for this insightful observation. We fully agree that different variables demand tailored diagnostic strategies. When developing MFM, our goal was not to replace these variable-specific diagnostics. Instead, we aimed to provide a general-purpose, robust metric that delivers a reliable overall score across accuracy, variability, and distribution similarity. These three dimensions have been

emphasized in numerous metric studies (Fu and Zhang, 2024; Kling et al., 2012; Pool et al., 2018). We envision MFM and conventional metrics as complementary. For instance, in Fig. 4a, NSE and KGE successfully detected a single outlier while reporting a catastrophic failure, even though the overall simulation fidelity is high. Variable-specific metrics serve as powerful diagnostic tools for identifying particular error modes, while MFM provides a stable, comparable overall score. We have updated the manuscript to reflect this complementary perspective (P25L553-P25L558):

"Furthermore, the evaluation of different land surface variables often demands tailored diagnostic strategies. For instance, soil moisture evaluation typically emphasizes temporal correlation and unbiased RMSE, with mean bias playing a secondary role (Entekhabi et al., 2010), whereas highly seasonal variables such as evapotranspiration and leaf area index benefit from decomposition into seasonal cycles and anomaly components for more targeted diagnosis (Mahecha et al., 2010). MFM is not intended to replace these variable-specific diagnostic approaches; rather, we recommend reporting MFM alongside fit-for-purpose metrics to provide both a reliable overall score and detailed insight into specific error modes."

4. Compared with traditional error metrics, MFM involves more complex computations. Could the authors clarify the scenarios in which they recommend using this metric?

➔ Thank you very much. Based on the generality, robustness, and bounded nature of MFM, we recommend using MFM in the following scenarios:

(1) Multi-variable evaluation. When multiple land surface variables (e.g., runoff, soil moisture, latent heat flux) are evaluated simultaneously, their distributions and statistical characteristics differ substantially. Traditional metrics may behave inconsistently across these variables due to the sensitivity issues discussed in our synthetic cases. MFM provides a comparable score across all variables within a unified [0, 1] range.

(2) Model intercomparison and ranking. When a single reliable score is needed to rank competing models, MFM's robustness to error compensation and low-variability instability makes it more trustworthy than NSE or KGE for this purpose.

(3) Model development. When developers adjust parameters or add new physical processes, they risk accidentally degrading performance in other aspects. MFM can serve as a performance constraint to ensure that model updates do not reduce overall fidelity.

We have added the following text to the Discussion (P25L549-P25L553):

"Future LSM developments can include MFM as an overall performance score. Because it is strictly bounded within [0, 1] range and dimensionless, MFM is suitable for multi-model comparisons across different variables where variable-specific metrics fall short. MFM can also serve as an improvement constraint. To ensure that updates do not accidentally degrade overall performance, parameter calibration and coupling of new physical processes should at least maintain the MFM score."

5. In addition to evaluating the performance of estimated variables, error metrics are expected to help diagnose potential model deficiencies. While I recognize the advantages of MFM in some cases, how can its results be interpreted to identify specific problems in the model?

➔ Thank you for recognizing our work. MFM's diagnostic capability lies in its decomposable structure. Its three component dimensions reveal the specific nature of model deficiencies: A low ω with high PPF indicates large magnitude errors, while a low ω driven by low PPF points to timing misalignment as the primary issue; A low φ suggests that the model fails to capture the observed variability structure, for example, by underestimating the range of fluctuations; A low η indicates systematic distributional mismatch, such as a persistent bias that shifts the entire distribution.

We have added an explanation of Fig. 10 to provide clearer diagnostic guidance (P21L483-P21L485):

"MFM can be decomposed as shown in Fig. 10 to identify specific aspects of model failure: low ω indicates magnitude or timing errors, low φ suggests inadequate reproduction of observed variability, and low η reveals systematic distributional mismatch."

6. Eq (1): what are i and n?

➔ Thank you very much. We have added the description of i and n below Eq. (1) (P2L38):

"Subscript i denotes the i-th time step, n is the length of the time series."

7. Line 39: I do not think it is appropriate to refer to NSE as the standard metric for LSM evaluation, as this may be misleading. Although NSE is useful for normalizing model performance and enabling cross-basin and cross-model comparisons, it should not be considered inherently better than other metrics. Its application should be determined by the specific variable and purpose, and model errors are often best explained using multiple complementary metrics.

➔ Thank you for your excellent point. We agree and have revised the text to describe NSE as a "widely used" or "traditional" metric rather than "the standard" (P2L39-P2L40):

"Recognizing the limitations of scale-dependence of RMSE, Nash and Sutcliffe (1970) introduced the Nash-Sutcliffe Efficiency (NSE), which is widely used for LSM evaluation:"

8. Lines 42-44: It is precisely because of its quadratic formulation and high sensitivity to outliers that NSE is often used in streamflow evaluations with a particular focus on peak flows. Controversial conclusions are more likely the result of applying NSE in inappropriate contexts, rather than an inherent problem with NSE itself.

➔ Thank you for raising this important point. We fully agree with your statement. We have revised the manuscript as you suggested (P2L43-P2L45):

"The quadratic form makes it highly sensitive to outliers and thus well suited for streamflow evaluation with a focus on peak flow. However, the inappropriate application of NSE can lead to erroneous conclusions (Gupta et al., 2009; Legates and McCabe, 1999)."

9. Line 57: What are these limitations?

➔ Thank you for your careful reading. We have moved the sentence to the end of the paragraph and made the limitations explicit (P3L76-P3L80):

"These limitations, including error compensation, instability under low-variability conditions, sensitivity to outliers in non-normal and highly skewed distributions, and failure to reflect the true characteristics of the system, are not merely theoretical concerns but also lead to systematic biases in model selection, misleading performance rankings, and potentially incorrect conclusions about model skill across different land surface modelling regimes (Klotz et al., 2024; Knoben et al., 2025)."

10. Lines 59-61: The correlation term in KGE helps penalize this issue.

➔ Thank you very much. We have stated this distinction (P2L59-P3L61):

"Although the Pearson correlation coefficient r can help penalize this issue, relative variability α and bias ratio β can cause the KGE to assign higher scores to objectively more biased models."

11. Lines 64-67: This statement is rather vague. Could the authors provide a concrete example, for instance, specifying the data or variables involved?

➔ We thank the reviewer for spotting this issue. We have developed a synthetic analysis (Fig. S1 in the Supplement, attached below) to demonstrate the issue. In normal distributions (Fig. S1a), the mean, median, and mode converge due to symmetry. In skewed distributions (Fig. S1b), these three statistics diverge. For bimodal distributions common in highly seasonal variables (Fig. S1c), all three statistics fall between the two peaks and do not adequately represent either mode.

If we treat the skewed distribution as a simulation and the bimodal distribution as an observation, the bias ratio $\beta = 1.02$, suggesting near-perfect agreement. This arises because the mean compresses the entire distributional information into a single number, masking the fact that the simulation overemphasizes low values while missing the peak near 0.60. We have revised the manuscript to make this point clearer (P3L64-P3L65):

"These statistics are most effective for normally distributed data (Fig. S1)."

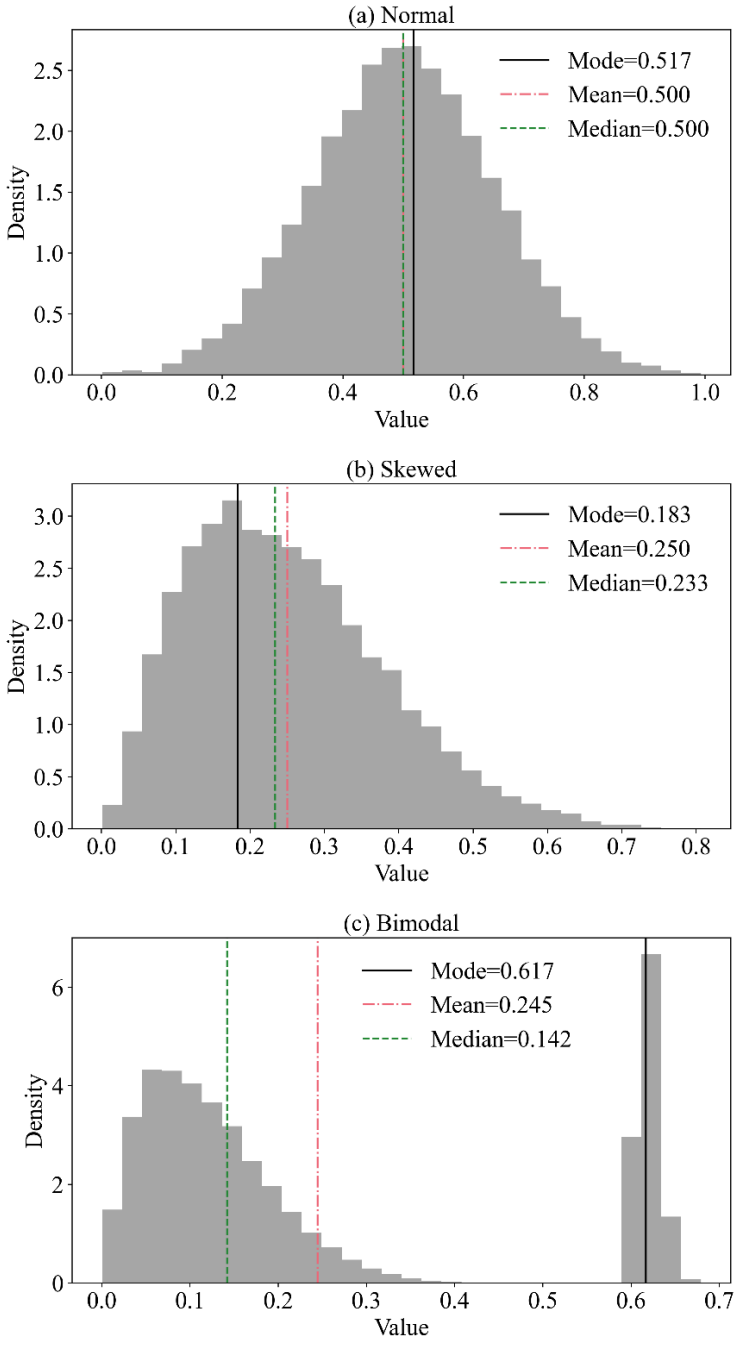

Figure S1. Artifacts of moment-based metrics. (a) Normal distribution. Mode, mean, and median converge. (b) Skewed distribution. Mean and median diverge from the mode. (c) Bimodal distribution. All three statistics become unrepresentative artifacts.

12. Line 69: Likewise, regarding the "right for the wrong reasons" issue, a concrete example would be helpful. This would allow readers to assess the severity of the problems potentially associated with KGE, rather than relying solely on the authors' statements.

➔ Many thanks for this important suggestion. We have added an example to illustrate this issue (P3L69-P3L71):

"For example, Cinkus et al. (2023) showed the bias ratio and variability scores were 11 % and 13 % respectively higher for the worse model. This occurs because the errors in worse model happen to cancel each other out, not because the model is genuinely better than models with lower KGE."

13. Line 71: If KGE is highly responsive to such balancing errors, what are the implications in practice? For instance, for simulations with similar KGE values, how large can the peak flow errors be?

➔ Thank you for your excellent question. To provide a quantitative example: as shown in Fig. 5a (Case 2, Scenario A), when observation variability is low, peak flow errors can grow arbitrarily large while KGE remains fixed at its initial value. This occurs because the error magnitude in the denominator of KGE's normalization stays small, making KGE insensitive to changes in the absolute error. In such cases, two simulations with identical KGE scores can differ by orders of magnitude in peak-flow error. In practice, we recommend two strategies to detect such issues: (1) examining whether regions with high KGE also exhibit high RMSE or NRMSE, which may signal error compensation; (2) using MFM alongside KGE to provide an overall evaluation. We have added further discussion about error mode (P18L448-P18450):

"Given that error modes arise from multiple sources (e.g., magnitude biases, phase shifts, and temporal dynamics), a comprehensive assessment combining error metrics, correlation coefficients, and spatial error pattern maps is helpful for reliable error diagnosis."

14. Line 103: What applications?

➔ Thank you very much. We have revised this (P4L108-P4L110):

"Despite this, it remains structurally sound and potentially useful for many applications such as water resource management where total volume or seasonal trends are more critical than time series accuracy (Liu et al., 2011; Magyar and Sambridge, 2023)."

15. Lines 106-110: The authors claim that highly skewed, non-Gaussian distributions violate the normality assumptions of moment-based metrics such as NSE and KGE, potentially biasing model evaluation. But do NSE and KGE actually require normally distributed data, or is this statement an overgeneralization?

➔ We thank the reviewer for pointing out this issue. We totally agree and have revised this statement (P4L113-P4L116, also check Response 11).

"When applied to non-normal distributions, moment-based statistics can produce misleading artifacts. In a normal distribution, the mean, median, and mode coincide due to symmetry, making the mean a faithful representation of central tendency. In skewed or bimodal distributions, however, these statistics diverge, and condensing the entire distribution into a single summary value entails substantial information loss (Fig. S1)."

16. Line 152: While developing metrics less sensitive to error compensation is a worthwhile goal, it is important to recognize that any aggregated metric will inevitably reflect a combination of different error types (e.g., random, systematic,

or phase errors). Complete elimination of error compensation within a single metric may therefore be unrealistic.

➔ Thank you for your profound observation. We fully agree and have revised "immunity" into "mitigation".

17. Line 157: What is p-Error? The definition of the NMAEp metric is presented too abruptly without a clear explanation.

➔ Thank you for pointing this out. The "p" in NMAEp refers to the exponent in the generalized L_p-norm. The parameter p thus provides a continuous control over the metric's sensitivity to errors without altering the range. The clear definitions are in Sect. 2. We have added a statement to make it clear (P6L163-P6L164):

"MFM integrates four fundamental aspects of model performance into these three dimensions, with the detailed definitions provided in Sect. 2:"

18. Line 160: Also, hard to understand what is SUSE and why it is workable for addressing KGE's shortcomings.

➔ Thank you for raising this point. The Scaled and Unscaled Shannon Entropy difference (SUSE) quantifies the similarity in information content between simulation and observation (Pechlivanidis et al., 2010, 2014). It computes the Shannon entropy difference under two binning schemes: a common-range scheme (scaled) that captures differences in data range, and an individual-range scheme (unscaled). The maximum of the two is taken to ensure that discrepancies in either aspect are detected.

KGE's variability term α compares standard deviations, which can be severely distorted by outliers in near constant data. For example, in Fig. 4b, a single anti-phase outlier causes $\alpha = 0.33$ (suggesting a poor match in variability), whereas SUSE yields $\varphi = 1.0$ (correctly recognizing that the overall variability structure is intact). Conversely, in Fig. 6a, a single extreme overestimation produces $\alpha = 11$ (a drastically inflated variability ratio), while $\varphi = 0.94$ (correctly indicating that the overall information content is largely preserved). By characterizing the entire

probability distribution rather than a single summary statistic, SUSE provides a more stable assessment of variability.

We have added the advantage of SUSE (P9L265-P9L268):

"In LSM data, moment-based statistics often suffer from excessive compression and are overly sensitive to outliers (Fig. S1). A single extreme value can cause the rejection of an otherwise good model, or conversely, a specific error pattern can make a poor model appear adequate. Information-based metrics provide a more comprehensive representation of overall distributional characteristics and are therefore less prone to such artifacts."

19. Line 234: This is not attributed by skewed data. This is a general artifact of aggregate error metrics sensitive to sign cancellation, which can occur with any distribution, including normal.

➔ Thank you for the precise point. We have revised this (P9L241):

"This ratio is susceptible to sign cancellation."

20. Line 244: What are min ($\boldsymbol{S}$, $\boldsymbol{O}$) and max ($\boldsymbol{S}$, $\boldsymbol{O}$)? $\boldsymbol{S}$ and $\boldsymbol{O}$ denote scaled and origin?

➔ Thank you very much. $\boldsymbol{S} = \{S_1, S_2, \cdots, S_n\}$ denotes simulations and $\boldsymbol{O} = \{O_1, O_2, \cdots, O_n\}$ denotes observations. Notation min($\boldsymbol{S}, \boldsymbol{O}$) denotes the minimum of $\boldsymbol{S}$ and $\boldsymbol{O}$, and max($\boldsymbol{S}, \boldsymbol{O}$) denotes their maximum. We have revised the text to make it clear (P9L252-P9L253):

"First, the scaled Shannon Entropy Difference (SED_{scaled}) is calculated by binning both time series using a common range from the minimum of simulations ($\boldsymbol{S}$) and observations ($\boldsymbol{O}$) to their maximum."

21. Line 290: Given that the limitations of NSE and KGE are discussed earlier, it is unclear why they are treated as benchmark metrics here. Would it be more appropriate to refer to them as baseline metrics?

➔ Thank you very much. We agree. Revised as you suggested.

22. Line 303: The introduction of the CAMELS dataset should appear earlier.

➔ Thanks for your careful reading. We have moved the introduction of CAMELS dataset to Sect. 3.2 (P12L324-P12L335).

Citation: https://doi.org/10.5194/egusphere-2025-6212-RC1

References

Cinkus, G., Mazzilli, N., Jourde, H., Wunsch, A., Liesch, T., Ravbar, N., Chen, Z., and Goldscheider, N.: When best is the enemy of good - critical evaluation of performance criteria in hydrological models, Hydrol. Earth Syst. Sci., 27, 2397–2411, https://doi.org/10.5194/hess-27-2397-2023, 2023.

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., Gharari, S., Freer, J. E., Whitfield, P. H., Shook, K. R., and Papalexiou, S. M.: The Abuse of Popular Performance Metrics in Hydrologic Modeling, Water Resour. Res., 57, https://doi.org/10.1029/2020WR029001, 2021.

Entekhabi, D., Reichle, R. H., Koster, R. D., and Crow, W. T.: Performance metrics for soil moisture retrievals and application requirements, J. Hydrometeor., 11, 832–840, https://doi.org/10.1175/2010JHM1223.1, 2010.

Fu, T. and Zhang, C.: Towards a generic model evaluation metric for non-normally distributed measurements in water quality and ecosystem models, Ecol. Inform., 80, https://doi.org/10.1016/j.ecoinf.2024.102470, 2024.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, J. Hydrol., 424, 264–277, https://doi.org/10.1016/j.jhydrol.2012.01.011, 2012.

Legates, D. and McCabe, G.: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35, 233–241, https://doi.org/10.1029/1998WR900018, 1999.

Liu, Y., Brown, J., Demargne, J., and Seo, D.: A wavelet-based approach to assessing timing errors in hydrologic predictions, J. Hydrol., 397, 210–224, https://doi.org/10.1016/j.jhydrol.2010.11.040, 2011.

Magyar, J. and Sambridge, M.: Hydrological objective functions and ensemble averaging with the Wasserstein distance, Hydrol. Earth Syst. Sci., 27, 991–1010, https://doi.org/10.5194/hess-27-991-2023, 2023.

Mahecha, M. D., Reichstein, M., Jung, M., Seneviratne, S. I., Zaehle, S., Beer, C., Braakhekke, M. C., Carvalhais, N., Lange, H., Le Maire, G., and Moors, E.: Comparing observations and process-based simulations of biosphere-atmosphere exchanges on multiple timescales, J. Geophys. Res., 115, G02003, https://doi.org/10.1029/2009JG001016, 2010.

Mizukami, N., Rakovec, O., Newman, A., Clark, M., Wood, A., Gupta, H., and Kumar, R.: On the choice of calibration metrics for "high-flow" estimation using hydrologic models, Hydrol. Earth Syst. Sci., 23, 2601–2614, https://doi.org/10.5194/hess-23-2601-2019, 2019.

Pechlivanidis, I., Jackson, B., and Mcmillan, H.: The use of entropy as a model diagnostic in rainfall-runoff modelling, Int. Congr. Environ. Model. Softw., 2, 1780–1787, 2010.

Pechlivanidis, I. G., Jackson, B., McMillan, H., and Gupta, H.: Use of an entropy-based metric in multiobjective calibration to improve model performance, Water Resour. Res., 50, 8066–8083, https://doi.org/10.1002/2013WR014537, 2014.

Pool, S., Vis, M., and Seibert, J.: Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency, Hydrol. Sci. J., 63, 1941–1953, https://doi.org/10.1080/02626667.2018.1552002, 2018.