**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 $n_{\mathrm{SUSE}}$ and $n_{\mathrm{PHI}}$ 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 $\omega$ 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 $\omega$ with high PPF indicates large magnitude errors, while a low $\omega$ driven by low PPF points to timing misalignment as the primary issue; A low $\varphi$ suggests that the model fails to capture the observed variability structure, for example, by underestimating the range of fluctuations; A low $\eta$ 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 $\omega$ indicates magnitude or timing errors, low $\varphi$ suggests inadequate reproduction of observed variability, and low $\eta$ 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 $\alpha$ and bias ratio $\beta$ 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)."

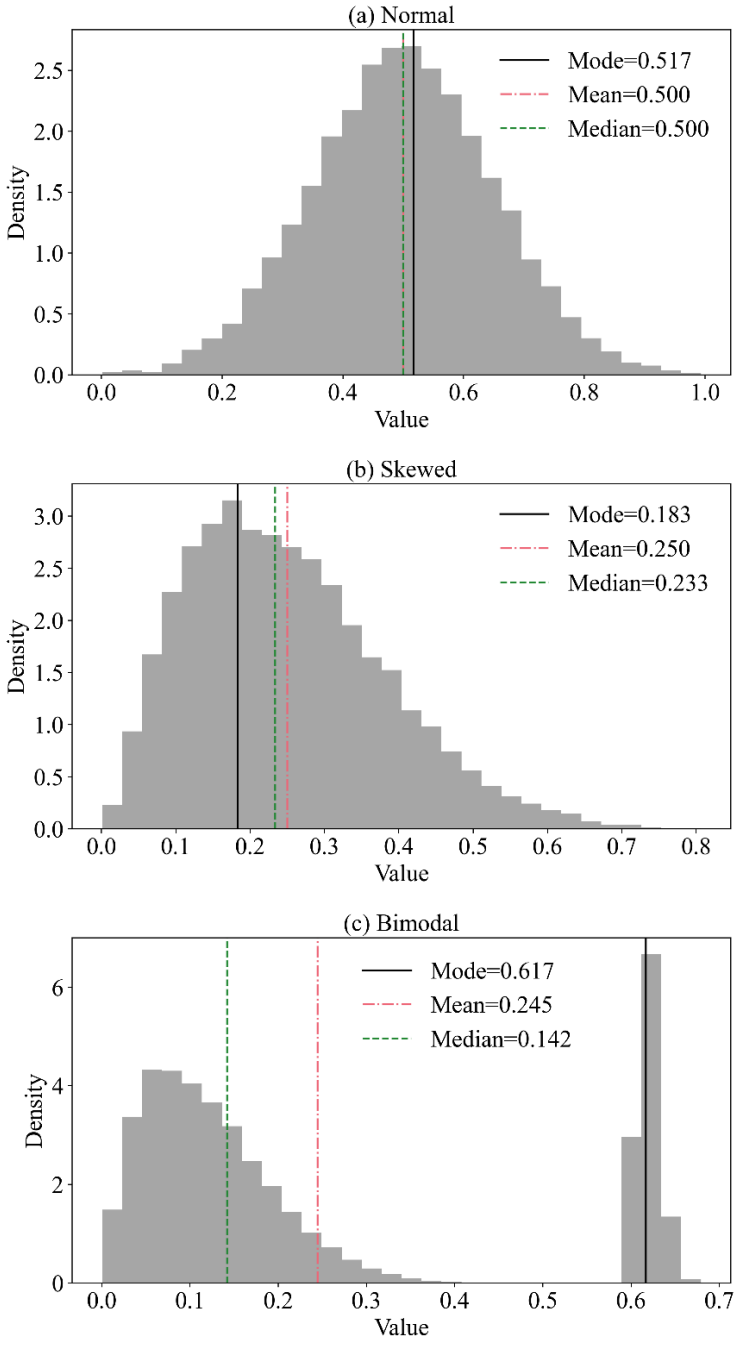

Figure S1. Artifacts of moment-based metrics. (a) Normal distribution. Mode, mean, and median converge. (b) Skewed distribution. Mean and median diverge from the mode. (c) Bimodal distribution. All three statistics become unrepresentative artifacts.

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 $\alpha$ 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.

**RESPONSE TO REVIEWER #2 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 #2 for thoughtful and constructive feedback. This Response to the Reviewer file provides 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. Quotations from the revised manuscript are shown in blue.

1. This study proposed the Model Fidelity Metric (MFM) framework which integrates three orthogonal dimensions of model performance within a Euclidean framework, including 1) Accuracy (NMAEp) penalized by a Phase Penalty Factor (PPF); 2) Variability (SUSE); and 3) Distribution Similarity (PHI). The synthetic cases and CAMELS data tests all showed that MFM provides a more authentic and reliable assessment of model fidelity compared to other traditional metrics (e.g., KGE, NSE, etc.). Overall, the manuscript is well written and provides a very useful LSM evaluation metric. I have a few comments and suggestions for the authors to address.

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

2. Major comments:

   1. The MFM framework depends on several tunable parameters, including the p value in NMAEp, the n_SUSE in SUSE, the c value in PPF, and n_PHI in PHI calculations. It is not clear how to effectively determine the values of these tunable parameters, which can affect the evaluation results especially when comparing with other traditional evaluation metrics.

➔ Thank you for this important suggestion. We have added specific parameter recommendations (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 $n_{\mathrm{SUSE}}$ and $n_{\mathrm{PHI}}$ 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."

3. 2. The proposed accuracy term tends to dampen the impact of phase mismatch error compared to traditional metrics (e.g., Case 3), which seems to make the phase error difficult to be revealed by just looking at the MFM metric value. For example, the anti-correlation in Case 3, where the model gets the pattern entirely wrong (reversed), should be evaluated as very poor performance, right? But when using MFM, the impact of this phase error is weakened relative to other metrics. This essentially does not penalize model much for its phase mismatch, which might mislead model improvements because the MFM value looks still OK (e.g., Case 3) even if the model gets the entire pattern wrong/reversed.

➔ Thank you for your insightful observation. This is a deliberate mathematical design intended to reflect the physical reality of simulation. MFM correctly captures this severe failure in the accuracy component ($\omega = 0.26$), where the PPF drops to its minimum (phase shift equals π and PPF = cos(π/c) = 0.707, c=4.0) and exp(-NMAEp) (p=1.0) drops to 0.36. However, the model still perfectly matches the distribution and variability ($\varphi = 1.00, \eta = 1.00$). This shows that while the timing is wrong, the model's physical capacity to simulate the correct data range (e.g., conservative law and variable dynamics) is still intact. We have revised the manuscript (P17L431-P17L433):

"MFM (0.572) rates the model as medium, where PPF (0.707) heavily reduces the accuracy score ($\omega = 0.260$). This medium value is caused by the variability and

distribution similarity components ($\varphi = \eta = 1.0$) recognizing the correct capture of the model."

While a single overall score is useful for evaluation, we agree it is inadequate for diagnosing specific and localized errors without decomposition. Users can decompose MFM as shown in Fig. 10 or use other metrics for indicating specific types of model failure. Furthermore, a score of 0.572 seems acceptable because MFM lacked standard benchmarks. Following your constructive suggestions and the recent guidelines (Clark et al., 2026), we defined performance thresholds for MFM and added them into the discussion (Figure S2 and S3 attached below) (P25L567-P26L588):

"To strengthen the interpretability of MFM and establish objective performance benchmarks (Clark et al., 2026), we conducted several calculations. We tested MFM against several naive predictive benchmarks under both default (Fig. S2) and enhanced (Fig. S3) parameter configurations. These benchmarks included the observation mean, uniformly and normally distributed random data, and percentile rankings. During these tests, the benchmark simulation sequences were sorted from smallest to largest to match the ranked observations. Additionally, we analyzed the metric's theoretical boundary conditions. If a model perfectly captures only one of the three MFM dimensions (e.g., $\omega = 1.0,\ \varphi = 0.0,\ \eta = 0.0$), it yields a baseline score of 0.183. If it perfectly captures two out of three dimensions (e.g., $\omega = 0.0,\ \varphi = 1.0,\ \eta = 1.0$), the score is 0.422. Furthermore, we derive the theoretical expectation of MFM for independent simulations ($s$) and observations ($o$) following an exponential distribution ($f(x) = \lambda e^{-\lambda x}$):

$$\mathrm{E[NMAEp]} = \frac{\int_0^\infty \int_0^\infty |s-o| \lambda^2 e^{-\lambda(s+o)} ds do}{\mu_O} = \frac{\frac{1}{\lambda}}{\frac{1}{\lambda}} = 1. \quad (17)$$

Where $\mathrm{E}[\cdot]$ is expectation. Under these conditions ($\omega = e^{-1},\ \varphi = \eta = 1.0$) MFM = 0.635. Analogously, the enhanced MFM calculates to 0.563 ($\omega = e^{-\sqrt{2}},\ \varphi = \eta = 1.0$). The exponential distribution was chosen because it is the most fundamental and easily solvable model for highly skewed LSM variables. Consequently, surpassing this threshold indicates that the model outperforms purely stochastic

systems which share the exact same statistical distribution. The superior threshold 0.8 was defined to filter out nearly all simple benchmarks, including mean, randomly distributed, and 10-bin benchmarks. As shown in Fig. S2 and S3, while 47 % of the 100-bin benchmark crossed the 0.8 line, only 2% of the enhanced setting reached this level.

To simplify usage, the benchmark thresholds are defined as 0.2, 0.4, 0.6, and 0.8. Scores within $[0.0, 0.2]$ are deemed unacceptable, indicating outliers. Range $(0.2, 0.4]$ is classified as poor, where the model equivalently captures at most one dimension correctly. Medium range is defined as $(0.4, 0.6]$, indicating the model equivalently captures at least two dimensions. Scores in $(0.6, 0.8]$ are considered as good, reflecting strong overall performance across accuracy, variability, and distribution. Finally, the $(0.8, 1.0]$ range represents superior performance, characterized by near perfect fidelity."

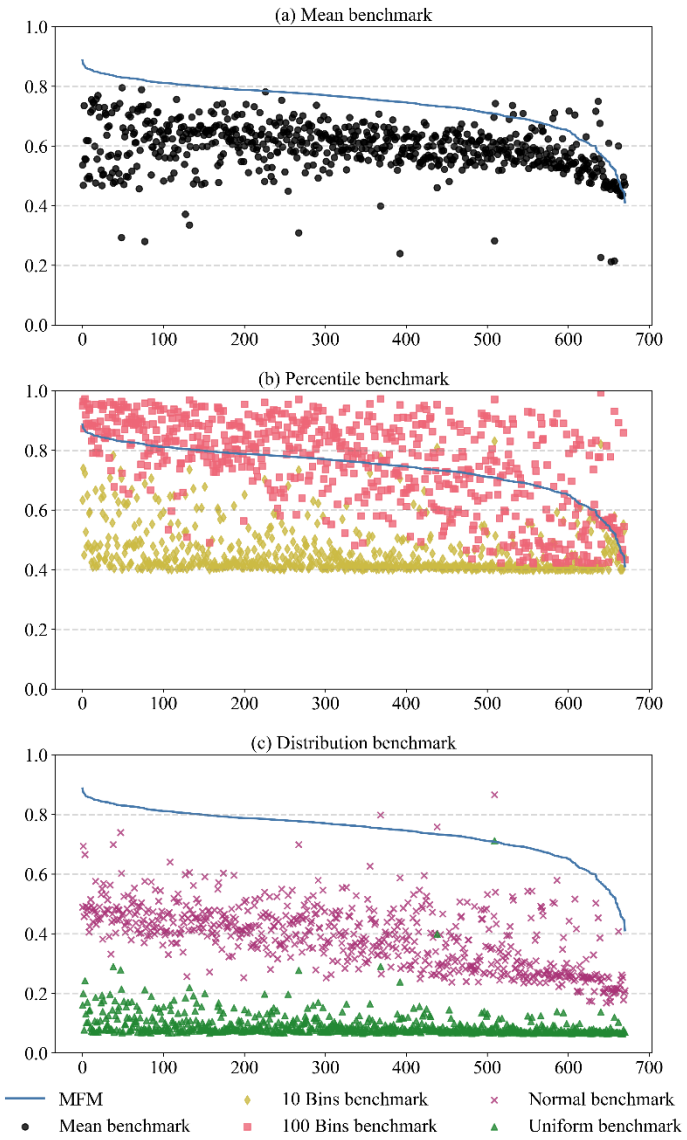


Figure S2. Benchmark evaluation of MFM (default parameters: $c = 4.0, p = 1.0, n_{\mathrm{SUSE}} = n_{\mathrm{PHI}} = 10$). Grey lines are 0.2, 0.4, 0.6, and 0.8. CAMELS observations are benchmarked against synthetic simulations paired in rank order. Each point represents a site. (a) Mean benchmark. Synthetic data equal the observed mean. (b) Percentile benchmark. Red squares denote 100-bin discretization (percentile) and yellow diamonds represent 10-bin discretization. (c) Distributional benchmark. Purple crosses indicate normal sampling (matching observed mean and standard deviation). Green triangles denote uniform sampling across the observed range.

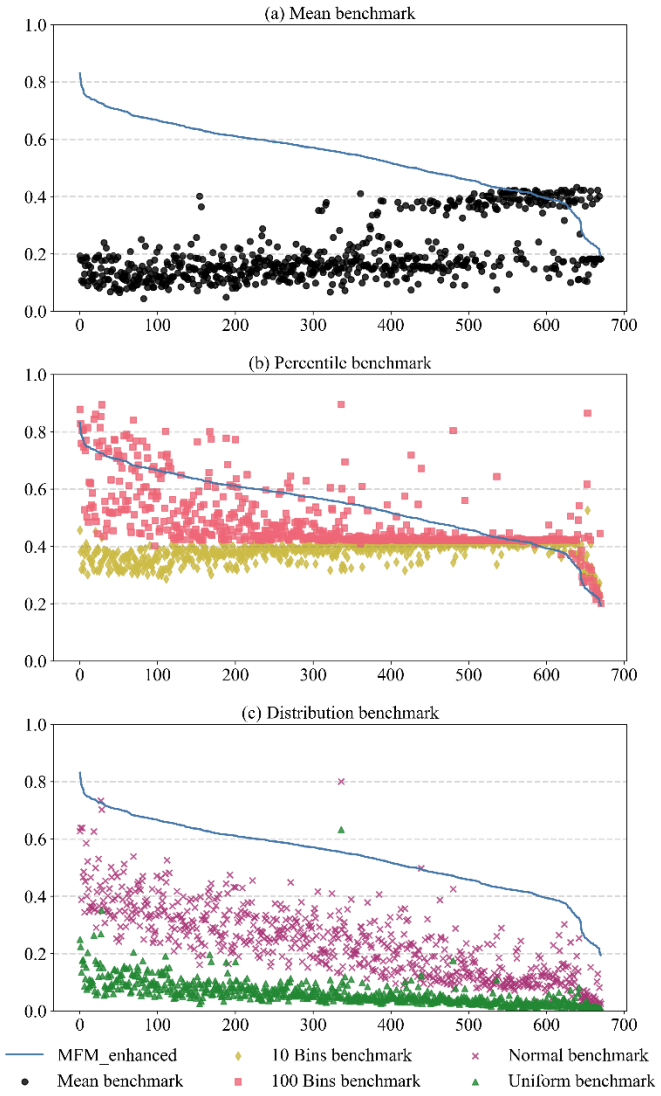


Figure S3. Benchmark evaluation of enhanced MFM (enhanced parameters: $c = p = 2.0, n_{\mathrm{SUSE}} = n_{\mathrm{PHI}} = 100$). Grey lines are 0.2, 0.4, 0.6, and 0.8. The methodology of generating synthetic simulations is the same as in Fig. S2. (a) Mean benchmark. (b) Percentile benchmark. (c) Distributional benchmark.

4. Minor comments:

   1. In the NMAEp calculation, how should users determine the p value, which seems also an arbitrary decision? Also, it is still not very clear to me why this formulation can avoid the impact of outliers when p is larger than 1. Should users compute NMAEp for multiple p values then?

➔ Thank you very much for your careful reading. We would recommend setting $p = 1.0$ for general evaluations, and increasing $p$ to 2.0 when strict accuracy is critical.

These exponents correspond directly to standard statistical metrics: $p = 1.0$ yields NMAE while $p = 2.0$ yields NRMSE.

You are correct about our assessment regarding outliers. Our previous phrasing in the manuscript is inaccurate. The $p$ value actually cannot avoid the impact of outliers. Setting $p = 1.0$ makes the metric less sensitive to outliers compared to higher $p$. We have revised the manuscript (P7L210-P7L211, check also Response 2):

"When $p = 1$, NMAEp is equivalent to the Normalized Mean Absolute Error (NMAE), which is less sensitive to outliers compared to higher $p$ values."

Generally, users do not need to compute MFM for multiple $p$ values. We have revised the manuscript to avoid ambiguity (P7L216-P7L218):

"Generally, we recommend using $p = 1.0$ for a balanced evaluation. When large error evaluation is critical (e.g., flood peak simulation), we recommend raising $p$ to 2.0."

5. 2. Equation (5): Is the d2 term within or outside the exponent p operator in Equation (4)? Is d1 here |Si-Oi|^p ?

➔ Thank you for pointing out this ambiguity in our notation. The $d_1$ and $d_2$ terms in Eq. (5) represent the absolute difference and do not include the exponent p. We have added statement (P7L220-P8L221):

"As illustrated in Fig. 1, an instantaneous error metric measures the vertical distance $d_1 = |S_i - O_i|$ where $|\cdot|$ means absolute value."

6. 3. Section 2.2.2: How to determine the number of bins (n_suse)?

➔ We thank the reviewer for this important question. Because land surface variables are always highly skewed, dividing the data into 100 bins inevitably creates empty bins and heavily crowded bins elsewhere. Therefore, setting $n_{\mathrm{SUSE}} = 10$ provides a more robust and statistically stable configuration. We recommend using the default setting ($n_{\mathrm{SUSE}}$=10) in general evaluation and increasing $n_{\mathrm{SUSE}}$ to 100 when

variability evaluation is critical. We have added statement for the choice (P9L262-P9L264):

"Setting $n_{\mathrm{SUSE}} = 10$ provides a more robust and statistically stable configuration. We recommend using the default setting ($n_{\mathrm{SUSE}} = 10$) in general evaluation and increasing $n_{\mathrm{SUSE}}$ to 100 when variability evaluation is critical."

7. 4. Equation (11): Why does it compute the minimal value of the two probabilities instead of their difference for each bin? The difference seems to better reflect their overlapping condition.

➔ Thank you for your excellent point. Your suggested approach and ours are actually mathematically equivalent. The overlapping ratio (the sum of the minimums, $\eta$) and half of their absolute difference add up to exactly 1.0. We chose to use the minimum simply because its formulation integrates seamlessly into our final $(1-\eta)^2$ formulation. We have added an explanation of this equivalence (P10L278-P10L280):

"It is worth noting that half of the absolute difference between probabilities is mathematically equivalent to $1 - \mathrm{PHI}$. We employ the minimum in this formulation to enhance the readability of MFM framework (see Eq. (15))."

8. 5. Section 2.3: I am a little confused here. Why does it need to use exponential transform for error and entropy components? There are other methods to normalize quantities to [0,1]. The exponential transform could introduce nonlinearity in the metrics which may favor more sensitivity to smaller values over larger values.

➔ Thank you for raising this important point. You are entirely correct. However, the exponential behaviour is a deliberate design:

(1) Handling unbounded upper limits. Error components (NMAEp) and entropy differences (SUSE) are unbounded on the upper end $[0, \infty)$. Standard linear normalization methods (e.g., min-max scaling) require a known absolute maximum value, which does not exist for general error metrics across different data.

(2) Representing marginal effects (Zhou et al., 2025). The nonlinear sensitivity mirrors the marginal effects of model improvement. The practical value of improving a model's error is not strictly linear.

(3) Matching the logarithm transform. The calculation of information entropy relies on a logarithmic transformation that maps data from a probability space to an information space. Therefore, an exponential transform naturally maps the information entropy back to the probability space [0, 1].

We have added statement (P10L285-P10L288):

"Additionally, the exponential transform introduces nonlinear sensitivity that prioritizes smaller errors over larger errors. This characteristic aligns with the marginal utility of model improvements (Zhou et al., 2025) while preventing extreme outliers from dominating the metric's dynamic range. The exponential transform is also a mathematical inverse of logarithmic operation in information entropy, mapping the entropy back into probability space within [0, 1]."

9. 6. Equation (12): The PPF is integrated into the accuracy term, which seems to make the decomposition and diagnosis of timing mismatch error difficult.

➔ Thank you for raising this practical diagnostic point. Your concern is practical and correct. The primary reason we did not separate PPF into another dimension is the behaviour of spectral phase analysis. Spectral phase shifts are heavily weighted by the dominant frequencies of the data (often the low variance, near constant background states). Evaluating phase shifts in isolation would miss the timing of short duration extreme events, leading to overestimation. However, if we just use extreme event timing error, the background states timing would be ignored, leading to underestimation. Hence, coupling overall temporal alignment (PPF) with errors (NMAEp) prevents the metric from overestimating while acknowledging the whole spectrum timing match. We have added statements (P10L292-P10L296):

"Spectral phase may be dominated by low variance background states. Evaluating phase shift in isolation overlooks timing errors in short duration extreme events, leading to overestimation, while focusing solely on extremes neglects the

background state, leading to underestimation. Coupling PPF with NMAEp prevents score inflation while ensuring comprehensive timing assessment across the entire spectrum. Therefore, we integrate PPF into the accuracy dimension rather than isolating it."

10. 7. I would suggest adding a discussion section to include: (1) guidance on how users can determine those tunable parameters in MFM framework in practice; (2) insights on how users can use MFM results to guide future LSM improvement.

➔ Thank you very much. We totally agree. Guidance on parameter selection has been added (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 $\omega$ to 0 under anti-phase conditions via PPF."

Practical insights on MFM usage have been added (P25L549-P25L558):

"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. 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."

Citation: https://doi.org/10.5194/egusphere-2025-6212-RC2

**References**

Clark, M. P., Knoben, W. J. M., Spieler, D., Gründemann, G. J., Thébault, C., Vásquez, N. A., Wood, A. W., Song, Y., Shen, C., Carney, S., and Werkhoven, K. van: Comment on Williams (2025): "Friends don't let friends use NSE or KGE for hydrologic model accuracy evaluation: A rant with data and suggestions for better practice", Environ. Model. Softw., 197, 106869, https://doi.org/10.1016/j.envsoft.2026.106869, 2026.

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.

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.

Zhou, X., Yamazaki, D., Revel, M., Zhao, G., and Modi, P.: Benchmark Framework for Global River Models, J. Adv. Model. Earth Syst., 17, e2024MS004379, https://doi.org/10.1029/2024MS004379, 2025.