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_{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."

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."

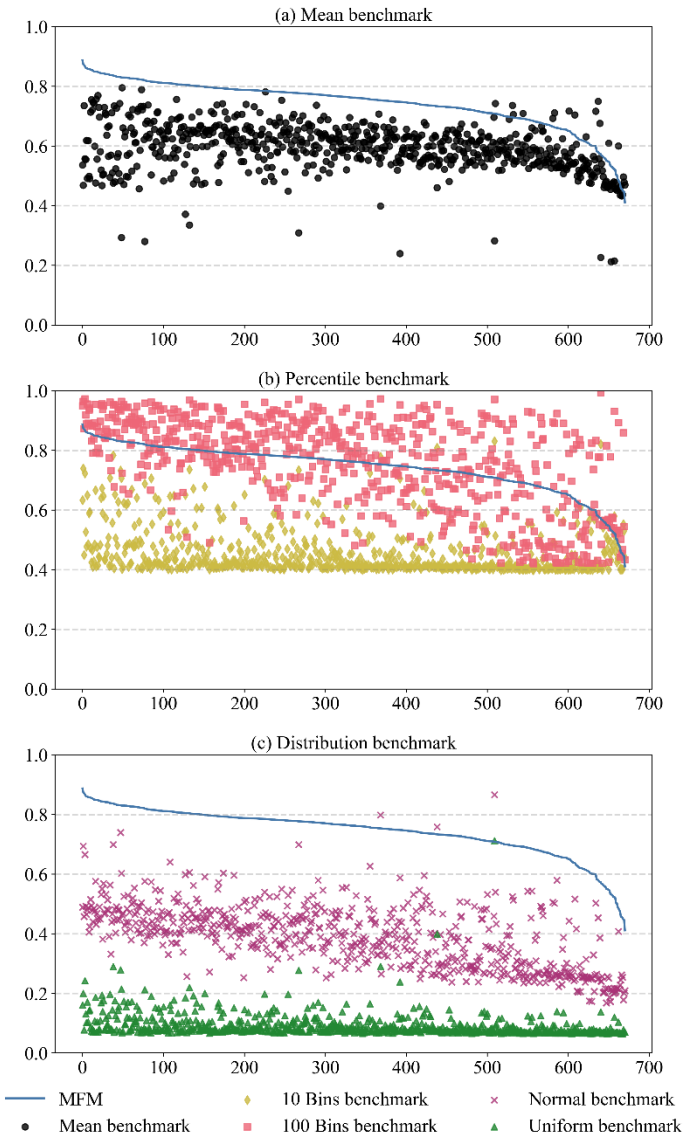

Figure S2. Benchmark evaluation of MFM (default parameters: $c = 4.0, p = 1.0, n_{\mathrm{SUSE}} = n_{\mathrm{PHI}} = 10$). Grey lines are 0.2, 0.4, 0.6, and 0.8. CAMELS observations are benchmarked against synthetic simulations paired in rank order. Each point represents a site. (a) Mean benchmark. 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.

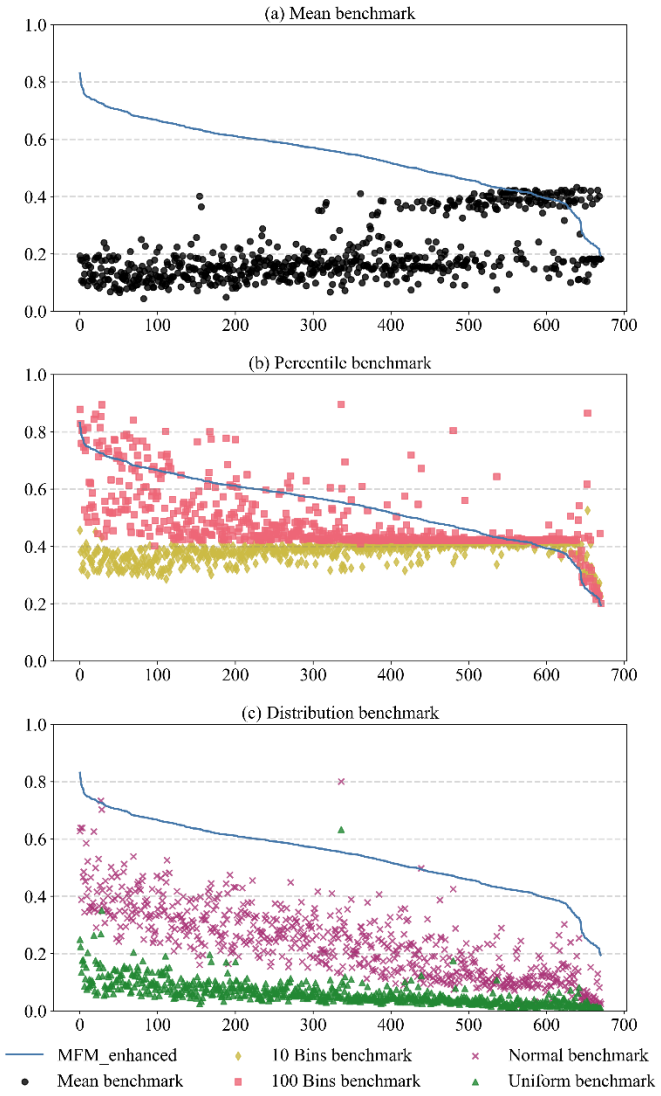

Figure S3. Benchmark evaluation of enhanced MFM (enhanced parameters: $c = p = 2.0, n_{\mathrm{SUSE}} = n_{\mathrm{PHI}} = 100$). Grey lines are 0.2, 0.4, 0.6, and 0.8. The methodology of generating synthetic simulations is the same as in Fig. S2. (a) Mean benchmark. (b) Percentile benchmark. (c) Distributional 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_{SUSE}=10) in general evaluation and increasing n_{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_{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, η) 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 ω 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.