Author's response to reviewers

We thank both reviewers for their thoughtful reviews and additional comments, which have further improved the clarity and readability of our manuscript.

We display the reviewers' comments in black and italic and highlight our responses in green and changes we made in blue in order to ensure clarity.

Best regards, Zavud Baghirov on behalf of the authors

1 Reviewer 1

Dear Authors,

Thank you for the careful and thorough revision of the manuscript. You have addressed the previous comments thoughtfully, and the manuscript has improved substantially in clarity, structure, and overall scientific rigor. After consideration of the following minor points, I would be pleased to recommend the manuscript for publication.

Dear reviewer 1,

We appreciate the additional points and suggestions you provided, which further enhance our manuscript's clarity.

1.1 Clarification of "data constraints"

Please clarify explicitly in the manuscript that the "data constraints" used to guide parameter predictions are "data-driven constraints" depending on model performance and generalization behavior. It would be important to state clearly that these are not physically defined hard bounds, but constraints emerging from the data and learning framework.

We thank the reviewer 1 for this helpful suggestion.

In the datasets section we now mention that:

"Throughout, we use "data constraints" to denote data-driven constraints implemented as observation-based loss terms using TWS, SWE, ET, runoff, fAPAR, GPP, and NEE. These terms guide the model's predictions but are not physically defined hard bounds or conservation laws. Physical constraints (e.g., mass balance) are enforced by the process-oriented component of the model."

1.2 Temporal and spatial resolution

The daily temporal resolution and the 1° spatial resolution (e.g., for WUE) should be explicitly stated in the manuscript. This information is essential for readers to properly contextualize the results and assess the scale of applicability.

We appreciate the reviewer 1's comment. We added the following sentence in the manuscript where we describe the model:

"Note that H2CM's static module outputs are at 1° spatial resolution, whereas the dynamic and process-based modules' simulations are at 1° spatial and daily temporal resolution."

1.3 Temporal evaluation metrics in the main text

The additional experiment and evaluation for the period 2018-2019 strengthen the manuscript. However, the corresponding evaluation metrics are currently presented only in the appendix figure. I recommend including at least one key quantitative metric directly in the main text to make the temporal evaluation more visible and accessible to readers.

We thank the reviewer 1 for this suggestion. We added the following text to the revised manuscript:

"For example, when testing our model on unseen years for monthly GPP and NEE values, Pearson's r is close to 1 for GPP and 0.96 for NEE. These results hold for both the spatial split (original model) and spatio-temporal split experiments, with very similar ranges across cross-validation folds."

With these minor clarifications and additions, the manuscript will be well suited for publication.

2 Reviewer 2

Baghirov et al. have made noticeable efforts to address the concerns raised by me (and reviewer 1). In general, the revised manuscript is clearer than before, and I understand better now what the authors have done. I appreciate the authors'

professionalism in addressing most concerns to some extent instead of dodging questions. Having said that, I believe that the manuscript can benefit more from further improvements in the presentation of the results. None of my points question the foundations of the scientific approach or the interpretation of results, but I think that the changes would substantially enhance the study's readability and therefore its success.

Dear reviewer 2,

Thank you very much for taking the time to review our manuscript and for providing valuable feedback to help us improve its clarity.

2.1 General suggestions

Framing: I understand now that the goal is to generate a reanalysis of carbon and water fluxes, and not to build a new land model (Sect. 3.4.1, page 22; and authors' replies to my questions). This makes sense. However, the title, abstract, and many other parts throughout the paper are in contradiction with this framing (or can at least be misunderstood). The abstract has virtually not changed since the last version and still starts with "We present the ... H2CM – a global model..." which is not wrong, but can be misleading. At least I initially expected a numerical model with time stepping schemes etc here. It would help if the abstract would state explicitly that the aim is to generate a flux reanalysis, using a combination of ML models, constrained by four algebraic linear equations linking T, GPP and NPP, and a simple equation for heterotrophic respiration Rh.

We thank the reviewer 2 for these suggestions. We added the following sentences in the abstract and introduction accordingly:

"... H2CM provides a"reanalysis" of recent land water-carbon cycle variations by combining multiple observational constraints synergistically..."

"... Thus, H2CM aims at facilitating a"reanalysis" of carbon-water cycle variations by a joint interpretation of diverse observational data streams based on conceptual process-understanding incorporated into deep learning."

I also still wonder what the model is really made for if it's not predictions. The model is trained on observations and it is shown that it can match these observations, but why is the "reanalysis" it offers better than what we already have in the observational datasets? My impression from the figures is mostly that H2CM follows the training data. I suspect that the argument is that one can recover certain variables, features or regions that are undersampled in observations, because the model is able to transfer its skill to other grid cells or features, as the authors show. This purpose of the model, and the evidence for its usefulness, should be worked out more clearly throughout the paper.

We thank the reviewer for this comment. We have dedicated an entire section to the reviewer's point ("Scope and strength"), which we have further revised to enhance clarity on these points.

2.1.1 Structure of the sections:

I still find the structure confusing. Very high-level and general information is provided rather late in the paper, while a lot of details are provided first. For instance, as mentioned above, the aim of the paper is explicitly explained only in Sect. 3.4.1 (page 22) – very late in the paper. Instead of providing such context after the results, it would help to be aware of the scope from the start. Also, I was missing a clear overview scheme of the whole approach, until I saw Fig. A1. I suggest to replace Fig. 2 by the current Fig. A1. Showing geographical maps is not needed there.

We thank the reviewer for these suggestions. We explicitly mentioned the scope in the abstract and introduction (please see our answers above). We also replaced the figure 2 with figure A1 (previously).

I am also convinced that the methods section should not start with the datasets, but with a high-level overview of the model, referring to the aforementioned figure. The order that would be most accessible to me would be: general approach, process-based equations, training method and loss function, datasets. It would also help to put a brief overview paragraph in the introduction which explains the structure of the paper. The information in the new paragraphs on page 22 and page 27 (Appendix point H) is extremely helpful, but it should be provided much much earlier, before the results are presented.

We rearranged the sub-sections in the methods secion suggested by the reviewer.

2.1.2 Relevance of the chosen hybrid structure:

The authors argue that the structural choices are beneficial, e.g. for the transparancy and even causality of the model. While I generally agree, it is not obvious that the results are practically better than what could have been achieved by one large neural network directly translating from environmental variables to ecosystem fluxes. For example, line 512-513 state that the model's skill emerges from the synergy of process-based and ML components, but this is largely a claim and not a result. Also, the authors state that NEE is better than in Fluxcom and TRENDY. Is it because of the structural constraint of Eq. 1-4? Or just because the machine learning models applied here allow a better fit? The authors should reflect on this a bit more.

We thank Reviewer 2 for raising these interesting points. During model development, we initially tested a single large neural network designed to receive all inputs and predict all outputs. Although this architecture reproduced the observations very well, we employed an explainable AI method (specifically, Integrated Gradients) and found that the model was engaging in shortcut learning (e.g., relying on implausible input-output relationships). This led us to adopt the 'guided' architecture. Since the use of explainable AI methods is outside the scope of this study, we decided not to include these findings in the current manuscript.

Regarding what contributes to the performance of H2CM, we added the following text:

"… Clearly, a key factor for the improved performance of H2CM compared to FLUXCOM in capturing NEE variations inferred from atmospheric inversions is the integration of this data constraint in the hybrid modeling approach. While this might also explain H2CM's improved performance compared to TRENDY, the machine learning component used to parameterize the conceptual process equations seems to contribute to the success significantly."

2.1.3 Performance and Evaluation

line 291-294: I don't understand the argument here. Why should a lower variance reduce correlations? It should be the opposite: variance affects the rmse, but does not affect the correlation. Also, what is meant with "long-term monthly anomaly"? I don't see any short vs long-term results in the figures.

We removed the sentence that caused confusion about variance and RMSE and replaced "long-term monthly anomaly" with "NEE monthly anomalies based on CarboScope data" for clarity.

Fig. 3: The monthly absolute data seems to perform almost identically to the mean seasonal data. One could remove the MSC bars and just say so, but I have no strong opinion about that. The fact that the mean seasonal cycle is captured with correlations very close to 1 implies that the phase is well caputured, I guess. The SDR is also close to 1, which to me implies that the amplitude is also well captured (at least for GPP and NEE (OCO-2). However, the rmse for GPP is very large compared to the deseasoned anomalies, which perform much worse regarding correlation and SDR. This is something I do not understand and which may need explaining. Lines 302-305 say that high RMSE "reflects errors in reproducing both the amplitude and phase of the seasonal cycle" – but why are Pearson's r and SDR almost perfectly at 1 then? Could it be that the annual cycle is so huge compared to the monthly anomalies that even a tiny relative error produces a comparably large rmse? Is that plausible?

We apologise for the confusion. We updated the paragraph for clarity:

"In terms of RMSE, H2CM tends to exhibit higher errors for monthly and seasonal data compared to monthly anomalies. This elevated RMSE primarily reflects the large magnitude

of the seasonal cycle, such that even relatively small absolute deviations lead to substantial RMSE values. Despite this, the near-unity Pearson's r and SDR indicate that the phase and relative amplitude of the seasonal cycle are well captured. In contrast, the monthly anomalies exclude the dominant seasonal component, resulting in smaller RMSE due to their reduced variance, but typically lower correlations because only irregular year-to-year variations remain."

2.2 Details

Fig. 2 should be replaced by A1 which shows the same structure but much more clearly. What I still miss in the figure is a representation of the training process: Shouldn't information be passed backwards in order to optimise parameters? Currently, the Figure shows a one-way flow of information. Also, if one could indicate which steps happen on daily time steps and which on monthly steps (e.g. using colours of text of boxes), that would help further. Ideally, one could even refer to the datasets listed in Table 1 by colour coding.

We thank Reviewer 2 for these helpful suggestions. In response, we have replaced Figure 2 with the revised version previously shown as Figure A1. We also added dashed arrows to indicate the backward flow of information during training (i.e., backpropagation).

Regarding the comment on daily versus monthly time scales, H2CM operates entirely at a daily temporal resolution; thus, there is no distinction between daily and monthly processes in the model structure itself. However, most H2CM estimates are constrained using monthly observational data, because daily observations are generally not available.

While Table 1 is great to get an overview of the used datasets, it can still become clearer which variable is generated on which time step, and on which resolution, in the model itself. Lines 85-87 really help already, but imply that most data is remapped to 1°, but some data is on 1/30°. How can the model work with mixed resolutions?

Please also see our previous response regarding the temporal resolution of H2CM. Table 1 shows the original temporal and spatial resolutions of the forcing data and observational constraints not H2CM's. H2CM itself operates consistently at a daily temporal resolution and a spatial resolution of 1°.

I am still not sure about the precise meaning of "data constraints" in this paper. Do the authors mean to distinguish the data that is used to train the LSTMs and FC-NNs in Fig. A1 from the data that is comparable to the output from the process-based part of the model (NEE, ET, TWS, ...)? Then please say so.

We updated the paragraph where we mention data constraints for clarity:

"Throughout, we use "data constraints" to denote data-driven constraints implemented as observation-based loss terms using TWS, SWE, ET, runoff, fAPAR, GPP, and NEE. These terms guide the model's predictions but are not physically defined hard bounds or conservation laws. Physical constraints (e.g., mass balance) are enforced by the process-oriented component of the model."

line 126 and elsewhere: "the process-based component". To my understanding, the only process-based parts are equations 1-4 (and perhaps something in the soil water balance module that is not explained here). Can the authors please be more specific and refer to the equation(s) in each case (here Eq. 1)?

The process-based component of H2CM simply contains the equations 1-4 in this manuscript and equations A1-A24 in Baghirov et al. (2025).

line 135: I am still confused about the way CO2 fertilisation is implemented. Due to the linear relationship without offset (zero-order term), GPP goes to 0 when CO2 does, and it doubles when CO2 doubles (all else being equal). This does not appear to be realistic, and the value of beta_CO2 changes nothing about that relationship. If I understand the authors correctly (line 138-141), they claim that the overall fertilisation is mediated by the effect of CO2 on alpha_WUE appearing in the same equation. But stomatal response to CO2 is a different effect, so the equation seems to be confusing too different mechanisms. Moreover, CO2 is not even an explicit input to alpha_WUE, so I don't follow the argument here.

We thank the reviewer for this point. Our implementation mimics a linear response of GPP to CO2 fertilization. While this is the simplest way to represent it, a dedicated experiment showed that the learned sensitivity to CO2 fertilization (β_{CO_2}) is underconstrained (Appendix D), i.e. it cannot be determined from the available data constraints as discussed in section Appendix D. Thus, a more complex representation with more parameters to learn is not justified at this point.

We agree that, in a purely theoretical sense, Eq. (2) implies an unrealistic behavior in that GPP would approach zero if atmospheric CO2 were zero. However, in practice this situation does not arise, because the model is driven by observed atmospheric CO2 concentrations, which are always far from zero over the study period. We therefore do not encounter this regime in application. We also agree that Eq. (2) is a highly simplified and conceptual representation of GPP rather than a fully mechanistic formulation of CO2 fertilization. Regarding α_{WUE} CO2 is intentionally not included as a direct input to this term. During model development, we found that this simpler formulation led to more realistic carbon-flux trends overall. The role of α_{WUE} is therefore not to explicitly represent CO2 fertilization, but to capture environmentally driven variability in effective water-use efficiency, while the direct dependence on atmospheric CO2 is represented by the multiplicative CO2 term in Eq. (2). More generally, this simplified formulation is intentional, as H2CM is designed to combine conceptual process structure with data-driven flexibility, allowing part of the observed behavior to emerge through the machine-learning parameterization.

line 154: I doubt if "stateful" is the best word here.

We removed "(stateful)" from the sentence.

line 166 and line 179, similar problem: "fully connected" is unclear, and line 167 "a dynamic NN" is confusing since the architecture is static and since all NNs simulate dynamics. It would help to add a sentence somewhere that makes clear what "static" versus "dynamic" means and find a good and consistent terminology for the NNs that generate spatial fields versus the ones that generate dynamics in time (time series).

We have replaced Figure 2 and the corresponding subsection with Figure A1 and the text in Appendix A, which already clarifies this aspect in detail.

line 197: What is meant with "some of these... are directly constrained"? Are any variables indirectly constrained, because they inherit the improvement from the constrained variables (observables)?

Similar to our previous response, we believe that the revised version of this section clarifies this aspect more clearly.

line 211-214: Mention that the data folds are sets of different grid cells. One still has to kind of guess otherwise.

We added the following sentences:

"To evaluate the generalizability of H2CM, we use a 10-fold cross-validation (CV) setup. For this purpose, the spatial domain is divided into 10 folds composed of different grid cells. We then train 10 separate models, each time leaving out one fold of grid cells as the validation set and using the remaining folds for training."

Table 3: "emerging global patterns" is a strange and unhelpful title. There are no patterns in the table.

We changed the title with "Carbon-water cycle metrics used in this study, with abbreviations, definitions, and units"

line 256: "each batch". It is not explained anywhere what a batch is, and this word only appears twice.

We briefly introduced what a batch is, when we first use the term:

"Therefore, we first compute the spatial mean of both CarboScope and H2CM NEE monthly anomalies during training within each batch (subset of training samples processed together in one optimization step)."

Sect. 2.3.2, the loss function. I still don't understand from the paper how different data constraints are weighted. If they were all on the same time and space scales, I get it that they are equal. But does the global mean of the CarboScope data affect the loss term with equal weight as a single grid cell from any other dataset

(i.e. almost not at all), or with the weight of any other global dataset? How is this implemented?

We thank the reviewer for this important comment. The effective contribution of each constraint technically depends on the scale at which it is evaluated. The CarboScope constraint for example contributes as a spatially aggregated time-series term rather than as a full gridded field. We did not introduce additional weighting because, in practice, the training was stable and we did not observe evidence that individual loss terms dominated or competed strongly with one another.

line 303: It should be "anomalies", not "anomaly".

Corrected.

Fig. 4: Each colour bars applies to three figures but is squeezed into the side of one figure. I suggest to place the bars next to the figures.

We thank Reviewer 2 for this suggestion. We initially tested the layout proposed by the reviewer, with the color bars placed next to the figures. However, this arrangement compressed the maps and resulted in substantial unused space in the figure layout. To make more efficient use of the available space and preserve the map size, we chose to place the color bars within the side of the map panels.

All other geographical maps: Same problem, please place colour bars next to the maps.

Please see our response above.

line 373: "expectation" (singular)

Corrected.

line 374-375: The authors state that they have used TRENDY model data to constrain H2CM. This is confusing since I don't see TRENDY mentioned as training data in Table 1. I then rememberd that this is mentioned in line 103-107, but I did not understand what precisely the "soft constraint" is and how it is implemented. Isn't it a strength of H2CM to rely on observed data and structural relationships, and not DGVMs which are often quite biased and oversimplified?

We thank the reviewer for this comment. As highlighted in the manuscript, we refer to this as a "soft constraint" because it is applied only to the spatial pattern of mean annual CUE, whereas the temporal dynamics of CUE are not prescribed and must emerge entirely from the training of H2CM. Its purpose is to guide the model toward plausible magnitudes and spatial patterns of mean annual CUE, thereby helping to constrain the partitioning between autotrophic and heterotrophic respiration, for which mean annual CUE appears to be relatively consistent across other models.

Sect. 3.2: I wonder if the authors are actually too fair to TRENDY DGVMs by using the model ensemble median, which is probably closer to observations than

any randomly chosen model. Comparing this median to only one realisation of H2CM feels like an unfair comparison in favour of TRENDY. H2CM only has to be better than the average DGVM, I believe (and would also be faster).

We thank the reviewer for this thoughtful comment. In fact, we do compare the spread of H2CM, represented by the 10 cross-validation folds, with the spread of the TRENDY ensemble in the figures. Because the spread among TRENDY ensemble members is typically much larger than the spread across the 10 H2CM CV folds, the H2CM uncertainty band may visually appear much narrower, which could give the impression that only a single realization of H2CM is being compared against the full TRENDY ensemble.

line 437: "explains most of the variation in NEE" compared to which reference? OCO-2?

We updated the sentence to explicitly mention OCO-2 MIP product as a reference.

line 438: why is Fluxcom so bad here (R2=0.16)? And isn't H2CM trained on Fluxcom? Should it not inherit the bias?

H2CM is not trained on FLUXCOM NEE, but constrained against OCO-2 MIP for this variable. Therefore, the FLUXCOM performance shown here is not directly inherited by H2CM. A more detailed discussion of FLUXCOM's performance against observations is provided in Nelson et al. (2024) and is outside the scope of this manuscript.

line 441-443 "accurately reproduces..." and so on: please refer to the figures where one can see this.

Done.

line 458: "drier", not "more dry"

Corrected.

Fig. 7: which year does it show? The black point on the map (Fig. 7a) is impossible to see, make it red.

We updated the figure and its caption to change the color to red and include the year information.

line 508: insert "the" before "study by Lee"

Done.

line 518: Bayesian with capital B

Done.

line 544: What is meant here with "initialization"? The model does not have a time-stepping scheme.

We thank the reviewer for this comment. Here, "initialization" refers to the initial values of the neural network weights prior to training. These parameters are initialized at the start of the optimization and are subsequently updated to minimize the loss function. In addition, H2CM does operate at discrete daily time steps.

line 548 + following + line 691 + elsewhere: "% 100 ppm -1" or even "15%100ppm-1" looks confusing.

We thank the reviewer for this comment. We used the units in a manner consistent with the existing literature, in order to maintain comparability with the studies cited in our manuscript.

line 552: Could the fact that beta_CO2 is not identifiable be related to the linearity of Eq. 2 as discussed above? Perhaps any value can be chosen and is then corrected by trainging alpha_WUE.

Since both parameters are learned by neural networks, it is possible that they partially compensate for each other, which probably leads to equifinality or reduced identifiability in H2CM. This is one reason why we additionally impose priors on the global constants. We discuss this issue further in Section 3.4.2 (Uncertainties and limitations) of the manuscript.

Some more info on the parameter calibration method would help.

We have revised the Methods section and the corresponding figure in the manuscript to provide additional details on how H2CM is optimized. Please refer to Section 2.1.3 ("Overview of the hybrid architecture") of the revised manuscript.

line 654: I know it is often used, but I never understand what "end-to-end" actually means. Please be more specific.

We removed the term "end-to-end" and explicitly stated that all trainable components of the hybrid architecture are optimized jointly by propagating gradients through the full model from the loss function to the trainable parameters:

"During training, all components of the hybrid architecture are optimized jointly, including the static and dynamic neural subnetworks and the differentiable process-based water–carbon cycle model. Gradients are propagated through the full model from the loss function back to the trainable parameters, so that all trainable components are updated simultaneously."

all time series shown in the figures: "across 10 CV folds": does that mean that the time series show spatial averages, and are also averaged over 10 randomly sampled folds?

We thank the reviewer for this question. The H2CM time series shown in the Appendix figures are not single fold-averaged lines, but shaded areas. These shaded areas represent the range across the 10 cross-validation folds at each time step. Since each fold produces a full 20-year global prediction, the spread of the shading reflects the variability among the fold-specific estimates.

Fig. 3, C9, D1: mark the value 1 (or 0, depending on the figure) with a horizontal line.

Done.

Appendix E: This means that for each parameter, we obtain one value l, and add all l's to the loss L in Eq. 5?

Exactly.

line 693: I don't understand what is meant by "may partly reflect a strong nudging term", please rephrase and/or explain.

We added the following explanation:

"We note that the observed agreement between estimates and reference values may partly be driven by the strong influence of the prior (regularization) term in the loss function, which constrains the model and reduces its ability to deviate from the reference."

References

Baghirov, Zavud, Martin Jung, Markus Reichstein, Marco Körner, and Basil Kraft. 2025. "H2MV (V1.0): Global Physically Constrained Deep Learning Water Cycle Model with Vegetation." *Geoscientific Model Development* 18 (10): 2921–43. https://doi.org/10.5194/gmd-18-2921-2025.

Nelson, Jacob A, Sophia Walther, Fabian Gans, Basil Kraft, Ulrich Weber, Kimberly Novick, Nina Buchmann, et al. 2024. "X-BASE: The First Terrestrial Carbon and Water Flux Products from an Extended Data-Driven Scaling Framework, FLUXCOM-x." *Biogeosciences* 21 (22): 5079–5115.