Response to Reviewer Comments concerning HESS submission egusphere-2022-1177

Tom Kimpson Margarita Choulga Matthew Chantry

Gianpaolo Balsamo Souhail Boussetta Peter Dueben Tim Palmer

August 14, 2023

Author response to referee reports for the paper egusphere-2022-1177, entitled *"Deep Learning for Verification of Earth-System Parametrisation of Water Bodies"*.

We thank the referees for their reading, helpful criticism and suggestions towards the improvement of the manuscript.

We have addressed each point in turn below and the manuscript has been updated accordingly

We hope that this satisfies the request for changes necessary to proceed with the publication of the updated manuscript.

1 Reviewer 1

Essential comments

1. **Essential comment 1: mixing of different models, parameters and predictors**

   We agree with the general points here raised by the reviewer that the terminology used was insufficiently precise. This has now been corrected throughout the entire manuscript. We now make it clear that we are always comparing the results between two neural network models, not comparing NN predictions with e.g. FLake. If we update some input field to our NN and the NN prediction accuracy improves, this is good evidence that the updated field itself is more accurate and would enable a physical model like FLake to make more accurate predictions. We note that this work is a first attempt / investigation to explore the possibility of using these kind of machine learning methods for a global check of surface physiographic fields, and we will look to refine and further develop these methods in future.

   Taking some individual points:

   - **L.79, L82-87**. What was meant here is the distinction between is the new information informative enough to have visible changes, or it is just lots of work for no impact? This has now been properly phrased in the text. Whilst we did briefly discuss local degradation previously we have also now included a more explicit discussion. Additionally, we explicitly calculate the training noise by retraining VESPER multiple times. We now highlight the effect of the training noise in the updated manuscript. Having retrained the model multiple times, our conclusions are generally unchanged for all grid-point categories, with the exception of the Vegetation category which has significant training noise over a small number of grid cells making it difficult to draw meaningful conclusions. This is again discussed in the text and we thank the reviewer for raising this point for our attention.
   - **2.2 L88-89** As the point above, we have retrained VESPER multiple times to show that the training noise is generally smaller than the prediction changes due to the different input fields. We reword the text throughout to be clear that we are always comparing two NN models.
   - **2.3 L105-106** As above, terminology and text has been changed throughout the manuscript to be more precise
   - **2.4 L19** As discussed we have now trained multiple versions of the model to better quantify the training noise and our conclusions remain unchanged that we can use VESPER to (a) check that an updated field is closer to reality and (b) see if this updated field increases the accuracy of our NN model. Both of these points are relevant for the updated fields within a physical model like FLake.

2. **Essential comment 2: strange results and speculative explanations**

   As mentioned we now explicitly calculate the training noise by retraining VESPER multiple times. For the points in Northern Canada, the Toshka lakes and the Vegetation category our previously quoted changes are less than the training noise. Again we thank the reviewer for highlighting this to us. We have removed the discussion on Northern Canada and the Toshka lakes from the manuscript, whilst the effect of the noise on the Vegetation category is now discussed explicitly in the text. We emphasize that our main conclusion re the lake and glacier categories are unchanged by this retraining - the difference in the improvement due to the updated fields is much greater than the training noise.

3. **Essential comment 3: MODIS observations**

   A discussion on the quality of MODIS data has been added to the text.

4. **Essential comment 4: Errors**

   Throughout the work we use an absolute error i.e $|\text{LST predicted by VESPER} - \text{LST from MODIS}|$. This is now specified explicitly in the text. We have also explored the use of different error metric such as bias and RMSE, but our conclusions remain unchanged.

5. **Essential comment 5: training and evaluating periods**

   We have trained VESPER with different input years (i.e. 2018, 2019) and results were the same. For monthly data training we agree that more data is needed and this work can be considered as our first attempt to represent and evaluate monthly lakes - the updated lake fields themselves are only for a single 12 month period. We have reworded the text accordingly to make this concession that this is our first attempt to include monthly lakes. Additionally data preparation should be more detailed - it would be useful to consider only grid cells with constant cover over the training period.

6. **Essential comment 6: VESPER does not "beat" ERA5, it corrects ERA5**

   Fully agree, corrected in text accordingly

7. **Essential comment 7: technical names**

   We have updated the definitions of the aggregation techniques in the text. For the k-nearest neighbours algorithm we feel that this is a sufficiently well known technique, common in many ML texts that it is sufficient to specify the technique used and reference the specific Python library (RAPIDS) that we used.

8. **Essential comment 8: predictors**

   The different VESPER models do indeed have a different vector of predictors. This has now been specified explicitly in the updated manuscript, along with better definitions with units of the various input fields (see Tables 1-3)

9. **Essential comment 9: vegetation and glacier updates**

   The vegetation and glacier fields were updated in proportion to the change in the lake fraction. For instance if before the fraction of the cell which is lake = 0.75 and the fraction which is glacier = 0.25, and then after the update the fraction of the cell which is lake = 0.80, the new glacier field is 0.20. This is now discussed in the updated manuscript.

10. **Essential comment 10: significant figures**

    Agreed. Corrected throughout manuscript

11. **Essential comment 11: seasonal lake fraction changes and salinity.**

    This is an good point. For this work we are satisfied to consider the combined affect of monthly maps and salt lakes, since many of the locations we specifically highlight in the manuscript are saline lakes with large expected time variability in the surface water. Further work in this area is currently ongoing and we defer a more in depth, global study of saline lakes and monthly maps for the future.

12. **Essential comment 12: What are shadows on Fig. 9-10? Please explain.**

    Previously these shadows were confidence intervals. In the updated manuscript, with multiple VESPER trainings these shadows are the $\pm 1\sigma$ bounds. We have specified this in the Figure captions.

Other comments

All typos and editorial comments have been corrected in the updated manuscript.

2 Reviewer 2

Comments

The reviewer suggests a major overhaul of the structure of the manuscript. We agree that this is a very good suggestion and the text has been completely restructured as recommend. We now present the construction of VESPER much more thoroughly, detail the various input fields, and specify the differences between the various VESPER generations. Only then do we then go on to deploy VESPER on lake fields and discuss the results. Where possible we have made an effort to be more concise.

- **l.54** The terminology re parameters, model and physiography has been updated throughout the text. The tables have also been updated to describe the choice of variables, and the different VESPER models (see e.g. Tables 1-3 in update manuscript).
- **l. 123 ERA5** All required information has been added to the text.
- **l. 133 MODIS** All required information has been added to the text.
- **l. 146 and Figure 3** By 4km resolution, we were referring to the resolution at the equator. This has now been updated in the text. Re the number of points at high latitudes, this is a natural consequence of the MODIS orbit, see e.g. animation at `https://svs.gsfc.nasa.gov/3348`
- **l. 184** This change has been made and a new table (Table 3) added which specified each VESPER configuration.
- **Figure 4 and related text**

  The prediction error is now defined at the end of section 2.4. Whilst the performance of VESPER relative to ERA5 is encouraging, one aspect of this is that VESPER has been trained directly on MODIS data whereas ERA5 has not. For this work we take MODIS data as our source of truth - as far as VESPER is concerned the MODIS data is reality, whereas of course the MODIS data has its own errors and systematics. The question of if a deep learning model could be used for forecasting is a very interesting one, but slightly beyond the scope of this study - we are primarily interested in quickly evaluating the accuracy of the fields that get passed to a dynamical model.
- **Section 3 Results**

  This section has been restructured to just contain the lake results, rather than the VESPER configuration as requested.

  A short discussion on how the non-lake climate fields such as vegetation cover or orography are update in response to the update in the lake fields is not included at the end of Section 2.2.1 c.f. Aral sea. We have tried to condense this section, but generally there is lots to discuss and we do prefer to be thorough here. As suggested there is plenty of further discussion to be had on e.g. monthly lake maps, glaciers etc. but we defer this for a future study
- **Section 4 Discussion**

  We have added the referee's point about if VESPER and ECLand parametrisations would react similarly to changes in the input fields to the discussion. In short, this is a very interesting question. The use of a tool that was trained from ERA5 in model output of IFS requires the assumption that the statistical behaviour of the fields does not change from ERA to IFS as the ML model would otherwise be forced to extrapolate (which it will not be able to do). This is a fair assumption, but it would be interesting to investigate this quantitively in greater detail. We defer the investigation of this question to a future work and thank the referee for raising a good point.