

Comments from Referee #2

We thank the reviewer #2 for his/her thorough comments on our manuscript. Please find below a detailed feedback to individual comments and questions

MAJOR COMMENTS

1. All discovered equations are linear. This requires explicit justification. In the Western Mediterranean, NPP is regulated by mechanisms that are intrinsically non-linear, both in the predicted field and in the predictor variables. For instance, phytoplankton growth rates, nutrient dynamics, and the photosynthesis-irradiance relationship are not captured by linear equations. Furthermore, the predictor variables exhibit non-linear long-term evolution, both in the historical period and under the RCP scenarios considered. Therefore, the authors should discuss explicitly whether the linearity result reflects a genuine physical property of the system or an artefact of the annual-averaging preprocessing and the complexity penalties inherent to the equation discovery algorithm. I am not implying that linearity is an algorithmic artifact. Annual averaging acts as a low-pass filter that can smooth non-linear functional responses. Therefore, the authors should explicitly discuss whether the linearity result reflects a genuine emergent property of the aggregated system or an artifact of the preprocessing and/or the algorithm. SR algorithms often aggressively penalise non-linear terms (divisions, exponentials, logarithms) because they increase equation length. If the penalty were relaxed, do non-linear terms (e.g., interactions or quadratic terms) appear, even if later pruned?

Yes, this is an excellent comment, which pinpoints a key point of our manuscript. As shown in Table 1 of the manuscript, all equations selected for the 9 different validation schemes are linear. More generally, when we analyze in detail the random search results, we observe that this algorithm discovers almost always linear equations: on average for every validation scheme, only 2.06% of the 500 equations discovered during the random search are interpretable non linear equations (see Sect 3.2 for the space of equations considered).

To answer your question, yes linearity could stem from an artifact of the equation discovery algorithm and/or from the annual averaging. Even if annual averaging may have linearized the relationship between annual NPP and regressors, our scientific objective is to predict annual NPP with an interpretable equation. So even if theoretically a linear equation may appear too limited/simplistic, in practice it fits well both of our needs.

In the discussion of the revised manuscript, we added the above paragraph to detail the two potential causes for the linearity of the equation. Furthermore, at line 289 we replaced “However, the advantages of our approach is that we do not assume any linearity, and allow also for non-linear terms in the final equation.” by “However, the advantages of our approach is that we do not assume any linearity, and allow also for non-linear terms in the final equation. A linear relationship can be a good approximation for the temporal and spatial scales analysed here. This linearization can also come from the complexity criteria of symbolic regression that seems to prioritize linear terms even if non linear terms are possible. In our future work we will try to understand this bias toward linear equations.”

2. The authors find that the linear equation underestimates interannual NPP variance, which is relevant for the interpretation of the findings. It is therefore necessary to quantify the magnitude of this underestimation relative to the total interannual variance of NPP. Furthermore, the implications for the use of these equations as operational emulators for future NPP projections should be discussed, as interannual variability is critical for economic applications and the ecological systems.

We agree with the reviewer: the underestimation of interannual variance is one the main limitation of our work. For all graphs with predictions, we added a metric that quantifies this underestimation in interannual NPP variance. Specifically, we compare (using relative error in %) the detrended standard deviation of the prediction with the detrended standard deviation of the reference. Note that in practice, the detrending is performed with the same linear function as for the other already existing metric “detrended correlation (COR)”.

In **Figure 4** below, we display this additional metric called “detrended standard deviation relative error (SDRE)”. As expected, we observe that the predicted standard deviation underestimates the reference standard deviation by -43.63% for the train set, i.e. $\sigma_{\text{reference}} = \sigma_{\text{reference}} - 0.43 \cdot \sigma_{\text{reference}}$, and by -46.47% for the test set. In the revised manuscript, this additional metric is shown.

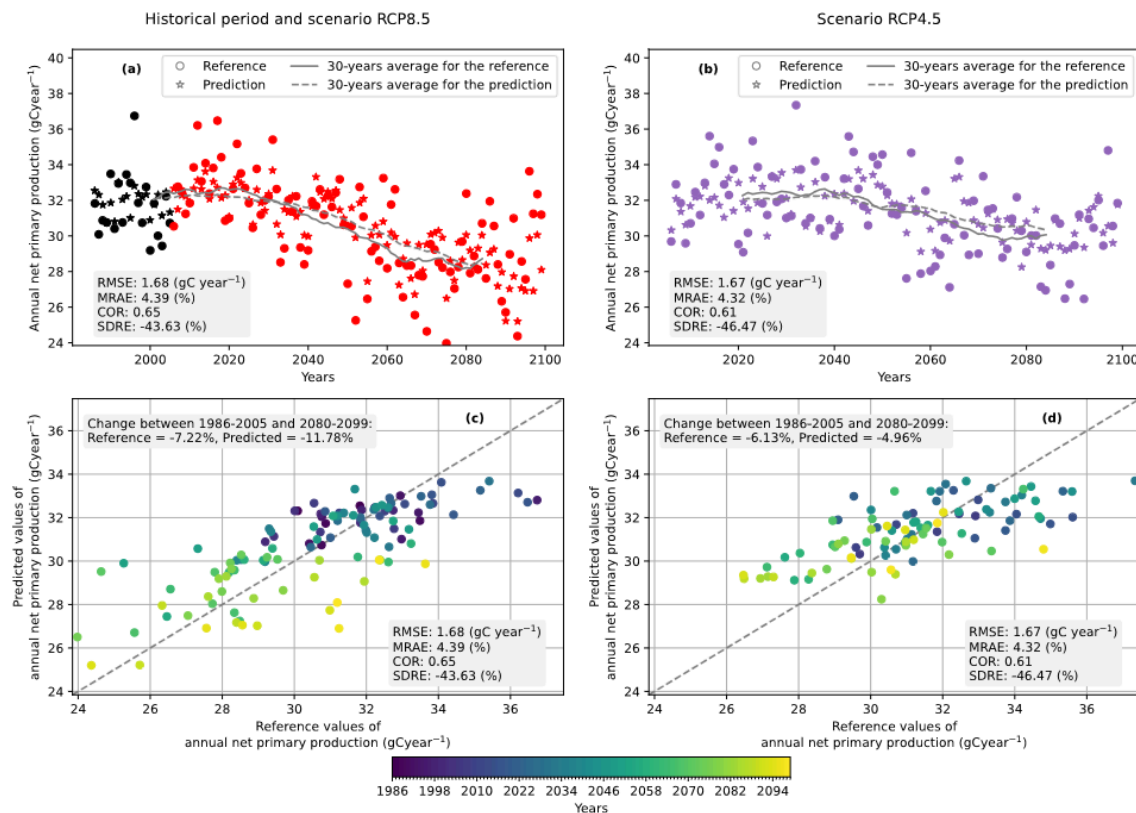


Figure 4. Comparison of predicted values by Eq. 2 with the reference values is illustrated with (a, b) time series plots and (c, d) scatter plots. This comparison is provided for two periods: (a, c) the historical period and scenario RCP8.5, and (b, d) the scenario RCP4.5. Several metrics are displayed: the root mean squared error (RMSE), the mean relative average error (MRAE), the detrended correlation (COR), the detrended standard deviation relative error (SDRE).

However, the underestimation of interannual variance is not necessarily inherent to the method. It is important to note that the “perceived” physics by the emulator is not the same as that perceived by the biogeochemical model, the former being low-frequency physical data represented by seasonal averages, while the latter consisted of daily current velocities and other variables provided every 3 hours, such as wind speed and shortwave irradiance. Consequently, given that the physical input data for the emulator are much more smoothed than the data used to force the biogeochemical model, and considering the model’s high nonlinearity and sensitivity to certain threshold effects, this underestimation of interannual variance is not really surprising.

To conclude, in practice, this fast emulator is reliable for mean predictions of the impact indicator (annual NPP) but not yet for interannual variability. If the objective is to evaluate ecological and/or economic risks linked to interannual variability, then this linear emulator is too limited. To improve predictions of interannual variability we could either add a stochastic term in the equation, or find additional predictors that match well the variance of annual NPP.

In the revised manuscript, we added a sentence in the conclusion underlying that this fast emulator is not suited for interannual variability predictions.

3. The systematic emergence of linear equations may reflect a structural bias of the methodology towards linear solutions, or it may reflect that the annual aggregation obscures scale-dependent non-linearities. A potentially valuable extension would be to apply a multi-scale decomposition approach such as Complementary Ensemble Empirical Mode Decomposition (CEEMD; Yeh et al., 2010) to both predictors and NPP prior to equation discovery. This could help isolate relationships that are individually linear at each timescale, even if the aggregate relationship is non-linear, and might better reproduce the interannual variance that the current linear equations underestimate.

This suggestion is really interesting, however we believe it is beyond the objective of this article. In the revised manuscript, we mentioned this suggestion in the discussion, as a perspective for future works. We note that such timescale decomposition would however lead to increased computational cost. For instance, we could estimate separately an equation for winter NPP, spring NPP, summer NPP, autumn NPP. However, such an approach would increase almost four times our computational budget, and would lead to a final equation for annual NPP (the average of the four seasonal equations) that would likely be far from concise. This is why instead we chose to directly find an equation that predicts annual NPP

MINOR

1. Change “As a perspective, this equation allows us to approximate the impact indicator at a neglected computational cost, i.e. without running the costly biogeochemical impact model, for any regional climate model outputs available” to “From this perspective, this equation allows us to approximate the impact indicator at a negligible computational cost, i.e., without running a costly biogeochemical impact model, using any available regional climate model output.”

Thank you for this comment. This sentence has been changed in the revised manuscript.

2. The absence of MLD from validated equations may reflect its thermodynamic definition, which captures buoyancy-driven stratification rather than the dynamical vertical mixing most directly relevant to nutrient supply. A more physically appropriate predictor for turbulent mixing could be the Ekman layer depth (h_E), which directly characterises wind-driven momentum penetration into the upper ocean. I suggest that the authors explicitly discuss this physical mismatch in the limitations section. If wind stress or wind speed data are readily extractable from the VHR-GCM and RCSM outputs used in this study, testing h_E (or a wind-based proxy) as an additional candidate predictor in a supplementary sensitivity analysis would be a valuable addition. However, if this requires substantial new data extraction and reprocessing, I do not consider it a strict prerequisite for acceptance; a thorough discussion of this limitation and a recommendation for future work would suffice.

Thanks for underlining this difference between the mixed layer, associated with the density anomaly definition in the model (buoyancy) and the mixing layer, associated with the turbulence scheme of the model (dynamic). We are well aware of this difference and both diagnostics exist in the NEMO model configuration used in the RCSM4 coupled model. At the daily scale, they can indeed differ over a few days when the mixing forces (buoyancy loss, wind stress) stop but the water column remains mixed before its restratification. However, this difference became neglectable when averaging over a year or a strongly-mixed season. After averaging over a season, the mixing layer diagnostic is generally slightly shallower than the mixed layer diagnostic but their year-to-year variability and their long-term trend are very similar. Therefore, we do not expect a different equation when changing the MLD definition. **Note that we are using the Mixed layer definition in the current study.** Note also that the mixed and mixing layer depth in DJF and in MAM in the studied area is much deeper than the Ekman layer due to ocean deep convection and therefore not relevant in our context. However, adding the Ekman layer depth would be very relevant in other seasons or regions of the world as wind may become the main MLD driver.

In the manuscript, we have completed the discussion on the obtained weak influence of the MLD variable. The current sentence (line 312 of the submitted manuscript) “Overall, we were surprised that the mixed layer depth (MLD), i.e. a variable that characterizes the vertical mixing, was not selected in any validated equations (Tab. 1). Thus, we could assume that here the SST and SSS indirectly integrate the MLD.” has been completed by “In this study, only one MLD definition has been tested (the mixed layer depth definition in the NEMOMED version). In the studied area (area of deep convection in winter), we do not expect significant changes in the results by changing the MLD definition. However, looking at the mixing layer depth or testing the Ekman layer depth as predictors could be very relevant in other seasons or regions of the world.”.