In my opinion the paper is worth to be published if a deeper discussion on results presented could be carried on.
Here a few suggestions:

Authors should go deeper in proposing founded reasons for models deviations from observed data.

Thanks for the proposed improvement. We will include a deeper discussion on the results. As an example, we will discuss some reasons why in Fig. 2 the time series of GPP for agriculture or evergreen broadleaved forest only show an r2<0.65.

Authors should put the validation of the model in a wider contest: do other models of the same type exist?

There exist other models of the same type as mentioned in the introduction (e.g. LPJ). We will discuss their results and compare with our results.

How do they perform when validated against similar databases?

The answer to this question will be included in our discussion (see previous question).

Is there consensus in the scientific community in this field on what a "good" model should look like?

No, there is no consensus in the scientific community what a “good” model should look like.

How does the presented model is positioned in this value scale?

It is difficult to answer this question, because there is no consensus in the scientific community what a “good” model should look like.

Related to the previous questions: usually models are developed for addressing a certain number of issues, as perfect correspondence between model results and measures on all parameters, whatever time and space scale is not usually feasible. It is not clear what is the key issue of the model proposed here: Figures 1 and 2 would suggest the spatial detail of the output is the key plus expected from the model, but in this case authors should focus their validation on systematically evaluating point-to-point correspondence with measured results, possibly using more advanced statistic performance indicators than the ones proposed in the paper. On the contrary, spatial and time averaged GPP values are presented in Tables 4 and 5, so giving the reader the impression the model had these values as main target.

One key issue of BETHY/DLR is indeed the spatial resolution of 1km x 1km. But a point-to-point validation is not possible neither using flux tower data nor other data. There is no validation data set available for GPP at a spatial resolution of 1km x 1km.
The spatial and time averaged GPP values shown in Table 4 are intended to show the yearly variability per country for the 8 years. And as demonstrated in Table 5 due to the

high spatial resolution of BETHY/DLR we can now link flux tower data representative for the CO2 exchange around the tower (10m – 100m) with 1km2 modeled data.

In conclusion, the work is interesting and in my opinion adds interesting material to scientific discussion in this field. Nevertheless model validation should be put in a wider context and more focused on proving the actual specific added value of the proposed model in comparison with state-of-the-art for the key parameters the model was designed to better reproduce.