PS: in the following weeks, we will also refine the language in the text.

In this revised MS, we have made the following modifications thanks to your suggestions.

1. Clarification of the terminologies

"**Model bias**" is used to denote the systematic errors of a land surface model, which should be distinguished from **ORCHIDEE error** (or **model-observations mismatch**) expressed as the sum of systematic error and random error from both observation and model. A brief definition of random and systematic errors is also introduced at the beginning of the "introduction". In the revised MS, we have emphasized that retrieved model bias (or **ANN-modeled error**) by ANN error modeling is only part of the 'full' model bias since some of the ANN inputs are mostly likely restricted by both data availability and our understanding of underlying processes. We have also stated that the main purpose of this MS is not to quantify the absolute values of model bias but to understand to what extent the model bias could be transferable in both spatial and timescale domains.

To make it clear, we used 'timescale' instead of 'frequency' in the text.

The « original » instead of un-decomposed time series is used to denote the time series without SSA decomposition.

2. ANN inputs used for ANN error modeling

To justify the ANN inputs (VPD, Rg and temperature), we have added the following into the revised MS.

"The ORCHIDEE errors from both original and reconstructed time series are then modeled by the ANN as a function of four drivers in their respective original and reconstructed forms. We should be in the mind that the main purpose of this study is not to quantify the absolute value of model bias but to understand model bias transferability in both spatial and timescale domains. Moreover, our choice of explanatory variables is constrained by the fact that a variable must be available at all sites. Based on this principle, we chose three meteorological drivers (T_a: air temperature, VPD: vapor pressure deficit, R_g: incoming shortwave radiation), and the simulated ORCHIDEE flux as ANN inputs. The three instantaneous meteorological inputs (T_a, VPD and R_g) were selected because they have been shown to explain the majority of variation in modeled instantaneous carbon and water fluxes on the one hand (Van Wijk and Bouten, 1999). On the other hand, these three variables have also been demonstrated to deliver the broad and universal nature of the model bias in a systematic sensitivity study, which is aimed to determine the optimal set of ANN inputs in carrying the greatest correction (Abramowitz et al., 2007). Other variables, such as diffuse PPFD, soil temperature and soil moisture are not considered since they are not available for all sites. In addition, the soil depth where soil temperature and moisture measurements occur is not uniform across sites."

For precipitation, its effects on terrestrial fluxes are mainly regulated by soil moisture, and although we could use decay rate of precipitation to approximate soil moisture

status, however, the coefficient characterizing this decay could be site-dependent, which is affected by soil texture, vegetation status and so on. The use of an index of soil moisture from an independent model or even ORCHIDEE itself could have been a solution to consider this. We did not test thoroughly any of these options, but one could try to get a more exhaustive set of predictor of model bias in the future studies.

3. Ensemble vs. The one with the largest RMSE-R

In the previous MS, we used 'the one with best performance (largest RMSE-R) in predicting the evaluation data', which could bias the result from the intuition.

To make things cleaner, we now use the mean of the ensemble. Note that the picture is almost the same as previously, and this is mainly attributed to our method of data splitting. "*We divide each variable (or node) in the input layer into different value ranges (or classes) according to their respective data distribution (e.g. Wijk and Bouten, 1999), and then build the training set in such a way that it would randomly select data from each combination of classes.*" In the beginning of this study, we did not build the data like this, and we have obtained notable differences between different ANNs.

4. Cross timescales analysis

The following sentence is added into the revised MS:

"In addition, all variables in both the input and output layers of the ANN have been rescaled into a [-1, 1] interval, and then the variable in the output layer has been scaled back to its original units before performing error calculations. In the cases where the variables in the training and evaluation datasets have the same order of magnitude and unit (e.g. error transferability investigation within a timescale), the variables in the two datasets are normalized together; otherwise, the variables in the two datasets are normalized separately (e.g. error transferability investigation across timescales, and between latent heat flux and GPP). This is done given the fact that an ANN error model trained on an output variable with an order of magnitude less than the one in the evaluation dataset (e.g. from interannual timescale to diurnal) can be problematic, as ANN are not robust extrapolation tools."

This choice of setup is very specific. We will present an alternative analysis with multi-linear regression as Appendic C, where we will consider the reduction obtained without any normalization for cross-timescales analysis (and with similar renormalization as the one done with ANN).

5 Site selection criterion

Our site selection allows data gap in meteorological forcing, but ORCHIDEE model could only be driven by the continuous climate forcing. To this end, we have added an appendix in revised MS, which documented the gap-filling algorithms for ORCHIDEE climate forcing for each site. In fact, when we perform ANN on the original time series, only time steps with actual measurements are used in ANN error model in order to avoid any bias due to gap filling methods by La Thuile data processing system (c.f. www.fluxdata.org). The reason why we did not use the data

tagged with high or medium quality because the gap in NEE or GPP is filled by either ANN or MDS (marginal distribution sampling, Moffat et al., 2007) method in La Thuile dataset
("*The ANN can be applied only if all the input data required are present. In case of gaps where one or more input was missing, only the MDS method has been applied.*"
c.f.http://www.fluxdata.org/DataInfo/Dataset%20Doc%20Lib/data_proc_scheme.pdf).

6 A note on Figure 3

We made a mistake on the Figure 3 in the pervious version, for example, the RMSE-R on interannual timescale is as high as ~90%. This extreme high value is induced by the fact that the training and evaluation datasets were inter-spread over time. Indeed, at the beginning of this study, to construct a training and an evaluation dataset for each site, we randomly select half of the data points from multi-year time series as a training dataset and the remaining one as an evaluation dataset. This way in splitting the data did not take into account strong temporal autocorrelations occurring in the dataset (e.g. Chevalier et al., 2011, Global Biogeochemical cycles, in press). To avoid this, we thus select the first half the time series as a training dataset and another half as an evaluation dataset.

7 PFT

We admitted that the vegetation type aggregated from Köppen-Geiger at each site documented in La Thuile could be problematic. But we could not find a favourable classification to replace it since more accurate one might not be available. We would like to believe that this will not drastically change the main conclusions of this study.

8 Some other changes

For ease understanding, we moved the introduction of SSA from Appendix to the methodology part.

We have rewritten the 'introduction and conclusions' thanks to your suggestions.