Dear Hyun Mee Kim,

Thank-you for the work that you and your co-author have undertaken to address the reviewers' comments on your manuscript. I have worked through your responses and I think some further work is required to justify the conclusions that you make in your paper or to determine whether they should be modified.

I have three main questions that I think you need to address.

1. Is the self-sensitivity really adding useful information to your network choice?

Apart from the methodological questions, about the usefulness of self-sensitivity, that reviewer 1 raised (Specific Comment 1 and 4), the evidence from your experiments for the improvement due to the self-sensitivity information appears to be relatively weak. In Figure 7 there is not much difference between the ADD and SS cases. In Figure 8 there is a larger difference between ADD and NSS, but the normalisation in NSS implicitly adds information about the ecoregion. I am wondering whether almost all of the improvement in your inversions is coming from the ecoregion information and not from the self-sensitivity. Even in the SS case, the restriction on your site selection (that all sites must be at least 1000 km apart) means that you are not solely dependent on the SS for your network selection, i.e. the network is more distributed over the Asian region than if you had only used the locations with largest SS values, which would have resulted in more sites in tropical Asia. The small impact of self-sensitivity is highlighted in the extra inversions that you performed in response to reviewer 2's request (Specific Comment 6) to consider additional current observing sites. When using the 18-site base case, all three cases with additional sites appear to give very similar results, with perhaps a very small improvement when ecoregion information is explicitly used in NECOSS1_18.

To try and confirm this, can I suggest another experiment which uses a network that accounts for ecoregion but ignores self-sensitivity (or actively avoids higher self-sensitivity)? This would use the same regions as ECOSS, but instead of using the maximum SS to choose the site(s) for each region, it would pick a random site for the region (or alternatively the lowest SS). If either or both of these experiments give a result that is comparable to ECOSS, then I think you will need to re-write the paper to say that you tested self-sensitivity as a method for choosing a network but that it didn't prove to be very useful compared to selecting based on ecoregion.

Incidentally, when creating your ecoregion based networks, why was region 145 excluded from ECOSS and additionally region 140 was not used for NECOSS1 and NECOSS2?

2. To what extent is your evaluation of networks dependent on the relationship between the hypothetical observations and your ecoregion-based inversion method?

Reviewer 1 (Specific Comment 2) asked about the impact of using hypothetical observations that are based on the same flux modules as used by the inversion, while reviewer 2 also asked a question (Specific Comment 3) about the hypothetical observations. By creating hypothetical observations based on changing the scaling factors for the fluxes rather than using an independent set of fluxes (e.g. from a different biosphere model), you are making the inversion easier by having the correct distribution of fluxes within an ecoregion. The inversion only has to correctly retrieve the scaling factor for a region and all fluxes within the region will be correct, whereas in reality we would not know that distribution precisely.

While the same hypothetical observations are used for all test cases and so the comparison between cases is valid, I still think there are important implications from the choice of hypothetical observations. For example, the improvement in the inversion from selecting sites by ecoregion must be influenced by using hypothetical fluxes based on ecoregions. If your inversion method did not solve for eco-region scaling factors and your hypothetical fluxes were not dependent on those scaling factors, then presumably there would be less value in an ecoregion-based network choice. Ideally, to assess this, you would need to do a new set of inversions that used a different set of fluxes (not dependent on the scaling factor) to generate hypothetical observations. Given that is probably not realistic as a revision to this paper, you need to be much clearer in your text about the dependence of your findings on the inversion method you are using and the choice of hypothetical observations.

3. Do all of your assessment metrics (PC, BIAS, RMSD) provide useful information and how are the metrics influenced by your inversion method and choice of hypothetical observations?

Reviewer 1 asked (Specific Comment 3) about the impact on your metrics from using 3 cases for REDIST and ADD compared to 1 case for your other tests and particularly about the bias as listed in Table 6 (Comment 34). In your response to the bias question you note that in the ADD case positive and negative biases tend to cancel spatially, and I assume, also across your 3 different ADD cases. For this reason, I do not think that BIAS is a good metric to include in your assessment as it hides compensating errors and tends to favour the ensemble tests (REDIST and ADD). I suggest you remove the bias panel from all figures and the associated text.

The RMSD seems to be your most reliable and least noisy metric and I recommend you focus your discussion on that metric. The spatial distributions in Figure 9 are useful for interpreting the timeseries figures and might be best introduced earlier. The spatial distributions also helpfully highlight the ecoregion dependence of the differences. In the timeseries figures I suggest that for the REDIST and ADD results you show a shaded band to indicate the minimum and maximum RMSD across your 3 cases, as well as the line for the mean RMSD. This would allow you to discuss whether other cases are better/worse than all three ADD cases or just better/worse than the mean.

I suggest you think about whether the correlation (PC) is needed for your assessment. PC is relatively high in all cases. I suspect this is due to the relationship between the inversion method and the hypothetical fluxes (i.e. with correct spatial distribution within ecoregions). Given the generally high PC, it is not clear to me whether there is any significance in the week to week variations in PC. Perhaps if you still want to include PC results in the paper, it would be sufficient to just keep them in Table 6 and only discuss their mean values across the experimental period.

I have a number of less important comments and suggestions but these are likely to be dependent on how the paper is revised based on the above questions, so I will not list these other comments here. Please let me know whether you are willing and able to undertake the major revisions that I think are required for the paper to be published. While the experiments you have performed are internally consistent and valid, I think that how they are analysed and interpreted needs to be reconsidered. In particular, I think you need to be careful to not make your conclusions too general when they may be quite specific to your inversion method and experimental design.

If you have any questions, please contact me by email (noting that I will be on leave through parts of January for our summer break so may be a little slow to respond).

Regards,

Rachel Law, rachel.law@csiro.au