In the present paper, the authors propose, test and apply a novel methodology to estimate changes in the global hydrological cycle (fresh water fluxes) from trends in ocean salinity. The method based on linear response theory and takes into account regional changes. The authors test their method utilizing ensemble simulation from a climate model (the Community Earth System Model; CESM) and apply it to estimate trends of the hydrological cycle from observations (temperature and salinity from the Institute of Atmospheric Physics). Applying the method to CESM data, show that the proposed methodology can reasonably recover the true freshwater flux of the ensemble mean. The generation of artificial ensembles allow for recovering also changes of individual realizations, i.e. indicate the applicability to observations. However, in this case additional significance criteria for the trend must to be met. Finally, the application to observations give results comparable to previous studies.

As the authors state, estimating the changes in the hydrological cycle as accurately as possible is important and a major challenge. Despite its limitations, response theory can provide an additional method to further improve the estimates.

Overall, I think this is an interesting and valuable study and provides sufficient new and significant information to warrant published. The manuscript is well written and structured and provides a (almost, see below) clear description of the methodology and the results. However, I have a few comments/questions (in random order) the authors may like to consider:

We thank the reviewer for their constructive comments, which we address individually below.

1) Surface temperature as an additional constraint: In lines 121/122 the authors state that they use surface temperature as an additional constraint. I'm wondering what effect this constraint has. How much would the results change with salinity only?

Thank you for the question. The surface temperature is a necessary constraint to help distinguish the response of regional surface salinity to different forcings, which gets blurred after dimensionality reduction. Especially in saltier regions, the response of salinity to heat fluxes and freshwater fluxes is similar, which is likely why the response to those two forcings can’t be separated based on surface salinity alone. The additional information from surface temperature constrains the problem by adding enough information that responses of the observables (regional salinity and temperature together) are sufficiently distinct between forcing experiments.

2) Response functions: From Figure 4 (and 5) the salinity responses of the three forcing experiments show some differences. Thus, the response functions (R) obtained from the individual simulations may have different properties as well. How are the actual Rs used for this study are related to the individual once (but, perhaps I missed or overlooked something)? In addition: As stated by the authors, the sixth mixture for the HadOM3 model seem to indicate a non-linear response. Are these data nevertheless contribute to the final Rs?

Thanks for the questions. As you note (from Figures 4 and 5), the responses due to the surface forcings differ between the FAFMIP ocean models that are used to form the response functions.

Here, our goal is to capture the individual response of regional salinity to different forcings when forming $R^h(t)$, $R^w(t)$, and $R^s(t)$ in a way that averages over model dependent responses. Thus, we use $R^h(t)$, $R^w(t)$, and $R^s(t)$ from each ocean model and carry through the procedure to find the response and then take the mean across ocean models at the end. We clarify this in text at line 181 by adding the following sentence: “We set-up Eq. (5) using response functions derived from each FAFMIP ocean model and then take the mean across ocean models (ACCESS-OM2, HadOM3, MITgcm) after solving for the $\frac{dF^h}{dt}$, $\frac{dF^w}{dt}$, and $\frac{dF^s}{dt}$ terms.”

As for the HadOM3 model, we do use it despite the non-linear response in the sixth mixture, as we wished to include as many ocean models as possible. However, we performed additional testing which demonstrates that the nonlinearity doesn’t impact the results significantly. We solve for the hydrological cycle amplification rate from observations again but ignore the sixth mixture when solving for fluxes from the HadOM3 model (the other two models continue to account for all 6 mixtures). We found a hydrological cycle amplification rate of 4.13 ± 1.21% which is well within the error bounds of the original result reported (4.52 ± 1.21%). The slight difference in results isn’t necessarily due to the non-linearity in HadOM3; rather the least squares problem solved is now differently constrained if we ignore the 6th region. However, the similarity in the results builds confidence in the robustness of the reported amplification rate despite the nonlinearity in HadOM3.

3) GMM regions and response functions: As far as I understand, the response functions are derived for the individual GMM regions based on CESM salinity (section 2.2.2). For the observations new GMM regions are defined (Figure 3). However, It seems that the response functions remain the same (based on CESM salinity). Is this the case? If so, how different would be the result when using response functions computed for the observation regions?

Thanks for the comment. For the observations, new response functions are used based on the GMM regions defined. We will clarify this in the text in the observations section. At line 268 in the original manuscript we add this sentence: “Following the steps from Section 3.1, we define the response functions $R^h(t)$, $R^w(t)$, and $R^s(t)$ from the response of salinity and temperature in the FAFMIP experiments in each of these GMM regions.”

4) Significance criteria: The authors need to apply additional significance criteria for the trends to capture the true response for individual CESM ensemble members. Unfortunately, the criteria (in my view) are quite subjective (or, better, are fitted to obtain the correct outcome for the given data set). Fortunately(?), the observations met the criteria for all regions. Beside that I'm surprised by this (which, in my view, may indicate important differences between observations and CESM data), I'm wondering how one would proceed in the case where not all regions met the criteria. Or: How large is the contribution of each GMM region to the total response?

Thanks for the comment. We agree that the criteria here are fitted based on testing on the CESM dataset. The motivation was assuming that different ensemble members of the CESM large ensemble sample enough range in the signal to noise ratio (forced response to internal variability) that we can derive reasonable significance criteria that hold on another dataset. In particular, for these significance criteria to hold across datasets, we needed ensemble members with lower signal to noise ratio than the new data that we wanted to apply the method to; this establishes a “worst case scenario” for which data the method can be applied to.

The fact that the observations easily meet the criteria suggests that observations have a stronger signal to noise ratio than most CESM ensemble members. If it had instead been the case that observations had not met the criteria, we wouldn’t have been able to apply the method to observations, as our method could not recover the true forced response due to strong internal variability.

As for the contribution of individual GMM regions, the fitted significance criteria indicate this to some extent: from testing on the CESM ensemble, having strong trends in regions 2 and 6 (which are the regions on average that tend to have the strongest response) is important for whether we can recover fluxes. In other regions, the trend tends to be unclear due to internal variability, but we can still recover fluxes despite this.

5) Effect of the heat flux: a) It seems that the authors relate changes in heat flux to changes in circulation/transport (e.g. L195) and therefore claim that their method also captures those changes (e.g. L. 332 & abstract). In general this may be true. However, since the heat flux (via the surface temperature) also directly affects evaporation and thus the hydrological cycle, it is not clear to me to what extent the effect of transport changes are really captured (using linear response). The authors may comment on this. In addition, evaporation also enters the heat flux (via the latent heat flux) and I'm wondering whether this matters for the derived (linear) response of salinity to heat fluxes (in particular for regions where evaporation dominates).

Thanks for the questions. Our method assumes that the response of the system (found through ocean salinity and temperature) to each of the FAFMIP forcings is separate. The ocean only FAFMIP experiments are performed by holding other fluxes constant while perturbing just one. For example, in the heat flux experiment, freshwater fluxes and wind stress are held constant while heat fluxes are changed, and so there can be no feedback between the perturbed heat fluxes and the hydrological cycle. Thus, any change in salinity in the FAFMIP heat flux experiment must be due to ocean transport change.

6) Figure 2: The authors may indicate the direction of the flux perturbation (does positive mean into the ocean?)

Thanks for the suggestion. We have added this to the figure caption. Here, positive indeed means downward, into the ocean.