Meteorologische Zeitschrift, PrePub DOI 10.1127/metz/2017/0790

Skill assessment of different ensemble generation schemes for retrospective predictions of surface freshwater fluxes on inter and multi-annual timescales

Vanya Romanova[1*], Andreas Hense[1], Sabrina Wahl[1], Sebastian Brune[2] and Johanna Baehr[2]

[1]Meteorological Institute, University of Bonn, Auf dem Hügel 20, 53121 Bonn, Germany
[2]Institute of Oceanography and Center for Earth System Research and Sustainability, University of Hamburg, Hamburg, Germany

(Manuscript received March 9, 2016; in revised form March 8, 2017; accepted April 23, 2017)

Abstract

The long term variability and its predictability of the monthly mean oceanic surface net freshwater fluxes is compared in a set of retrospective predictions. All are using the same model setup, and only differ in the implemented ocean initialisation method and ensemble generation method. The basic aim is to deduce the differences between the initialization/ensemble generation methods in view of the uncertainty of the verifying observational data sets. The analysis will give an approximation of the uncertainties of the net freshwater fluxes, which up to now appear to be one of the most uncertain products in observational data and model outputs. All ensemble generation methods are implemented into the MPI-ESM earth system model in the framework of the ongoing MiKlip project (www.fona-miklip.de). Hindcast experiments are initialised annually between 2000–2004, and from each start year 8 ensemble members are run for 10 years forward. Four different ensemble generation methods are compared: (i) a method based on the Anomaly Transform method in which the initial oceanic perturbations represent orthogonal and balanced anomaly structures in space and time and between the variables taken from a control run, (ii) one-day-lagged ocean states from the MPI-ESM-LR Baseline 1 system, (iii) one-day-lagged ocean and atmospheric states with preceding full-field nudging to re-analysis in the atmospheric and anomaly nudging in the oceanic component of the system – the Baseline MPI-ESM-LR system, (iv) an Ensemble Kalman Filter (EnKF) implemented into oceanic part of MPI-ESM, assimilating monthly subsurface oceanic temperature and salinity using the Parallel Data Assimilation Framework and full-field nudging in the atmosphere. The hindcasts are evaluated probabilistically using freshwater flux data set from NCEP-R2. On the global scale the physically motivated methods (i) and (iv) provide probabilistic hindcasts to some extent higher correlation and reliability than the lagged initialization methods (ii)/(iii) despite the large uncertainties in the verifying observations and in the simulations. We suggest similar approaches for further evaluations of other variables of decadal hindcasts systems.

Keywords: surface freshwater flux forecast, mean climate state statistics, probabilistic evaluation, ensemble generation methods

1 Introduction

Via their variability and the associated redistribution of energy and water, the large scale air-sea fluxes play an important role in climate variability on inter-annual to decadal timescales (Zaucker et al., 1994). The evaporation of water modifies the energy budget of the Earth and affects the climate (Trenberth and Guillemot, 1998). Additionally, the freshwater input by precipitation into the ocean at the ocean convective sites can alter the strength of the overturning circulation (Broecker, 1994) and redistribute the energy within the Earth climate system. Knowledge of the future variability of the oceanic surface freshwater fluxes is essential for an independent assessment of global climate variability and change.

Many studies concentrate on comparison of differently obtained freshwater products on annual, inter-annual and decadal timescales (e.g., Béranger et al., 1999; Adler et al., 2001; Stammer et al., 2004; Josey and Marsh, 2005; Romanova et al., 2010; Vinogradova and Ponte, 2013; Giglio and Roemmich, 2014; Iwasaki et al., 2014). The freshwater flux is a composite quantity, calculated from the difference of the precipitation (P) and evaporation or evapotranspiration (E). Each individual error in the two components contributes to the total error of the difference $P - E$ which could therefore be quite large. The errors depend on the methods used for producing the individual variable field. In-situ and satellite measurement errors are often systematic and the gridded fields show unbalanced conditions. Usually on a global mean the total freshwater fluxes does not sum to a small number. Differently from the in-situ/satellite estimates the atmospheric and ocean re-analyses provide model outputs for the freshwater fluxes. This should in

*Corresponding author: Vanya Romanova, Meteorological Institute, University of Bonn, Bonn, Germany, e-mail: romanova@uni-bonn.de

principle profit from conservation laws of the driving models, but due to model errors through discretization and parametrization the models of either atmosphere and ocean develop themselves systematic errors in their water and energy cycles. These errors are often compensated by non realistic fluxes from the other components e.g. the ocean provides infinite energy and water sources to an atmospheric model if there is no coupled model system. Partly, this can be fixed by the assimilation procedure used in the re-analysis, which constrains the model solution to the observed data and may provide more realistic and potentially balanced fields. Although most of the reanalyses systems assimilate very similar observational data sets, they differ somewhat from each other due to different assimilation techniques and dynamical model. As a result of the atmosphere-ocean dynamics, the net freshwater fluxes shifts the geographical locations of the source and sink zones. A correlation analysis hardly shows coefficients greater than 0.7 or only 50 % of common variance when compared to independent data (Romanova et al., 2010).

Recent years have seen different approaches for multi-year and decadal climate predictions (e.g., Keenlyside et al., 2008; Pohlmann et al., 2004; Smith et al., 2007; Kruschke et al., 2015). Especially the details of MiKlip decadal climate prediction system, the different approaches for data assimilation and ensemble generation are discussed in depth in Marotzke et al., 2016. One of the major issues in the prediction systems is the method for construction of the disturbances applied in the initial fields (Toth and Kalnay, 1993; Magnusson et al., 2008; Wei et al., 2008; Keller et al., 2010; Du et al., 2012; Haughton et al., 2014; Romanova and Hense, 2015). Its aim is to span the most probable range of the models phase space trajectories. The success of an ensemble generation scheme depends on the geographical location and space orientation of the initial disturbances, which will develop on annual or decadal time scales and will produce an adequate spread. Romanova and Hense, 2015 investigated the effect of orthogonal rotation on the disturbances patterns, different norms and re-scaling methods for generation of the disturbances. They argued that the orthogonal conditions and definition of the re-scaling factors are important to produce plausible spread, in contrast to the selected norm, which in their study was the total energy norm and ocean heat content norm. Studies on Singular Vectors (Molteni et al., 1996; Marini et al., 2016) and Ensemble Transform (Wei et al., 2008) as competitive ensemble generation methods also point to advantages of the orthogonal rotation of the perturbation patterns to achieve more skillful forecast.

Predicted quantities, which are mainly studied and analysed with respect to long term climate variability related to the ocean are the sea surface temperature, the Atlantic Meridional Overturning Circulation (AMOC), and the Ocean Heat Content (Matei et al., 2012). Moreover, the assessment of ensemble generation methods was mainly based on the evaluation of the predicted sea surface temperature with inconclusive results up to now. Additionally the uncertainties of the observational data sets or reanalyses are not taken into account. In contrast to previous attempts we concentrate in the following on the evaluation of the air-sea freshwater flux (P-E) and measure the predictability skill relative to NCEP R2 reanalysis data (Kalnay et al., 1996). This special choice is justified by a comparison with other available and comparable data sets. We concentrate our analysis only on one variable, since multivariate analysis could be complicated. When turning to a family of variables the multivariate character of verification has to be incorporated, taking into account the correlations among the variables. The freshwater flux is a variable which determines the temporal change of state variables like salinity in the ocean or water vapour concentration in the atmosphere. By this property it provides orthogonal or other information relative to the state variables. Additionally the freshwater flux has a direct physical interpretation in the sense that it is one of the coupling components within the hydrological cycle between the atmosphere and the ocean.

The paper is organized as follows: in the second section the model, the ensemble generation schemes and the hindcasts are described. The third section deals with the evaluation scores and the analysis and discussion of the results are given in the fourth section.

2 Ensemble generation methods and data

All hindcasts are performed using the MPI-ESM-LR coupled model developed at Max Planck Institute for Meteorology in Hamburg (Giorgetta et al., 2013). The hindcasts analyzed in the current study differ in the ensemble generation schemes. Anomaly Transform (Romanova and Hense, 2015) provides re-scaled orthogonal perturbations under TE (Total Energy) norm. We consider two types of lagged hindcasts: one is produced by applying one-day-lagged perturbations only in the ocean and the other uses synchronized perturbation in the ocean and atmosphere (Baseline 1). The fourth hindcast is based on EnKF filter (Brune et al., 2015). All hindcasts cover the same time period, using five starting years, from 2000 until 2004, with 8 or 10 ensemble members, each is run for 10 years forward.

2.1 Model and hindcasts

The MPI-ESM couples the following components: a) the ECHAM6 atmospheric spectral model (Stevens et al., 2013) at T63/L47 resolution corresponding to approximately 1.87° horizontal resolution on 47 vertical levels.; b) MPIOM ocean model (Jungclaus et al., 2013) on a bipolar curvilinear GR15/L40 grid with approximately 1.5° horizontal resolution on 40 non-equidistant vertical levels; c) the Sea-Ice component on the ocean grid; and d) the JSBACH land surface model. The components exchange energy, momentum, water and trace

gases via the OASIS coupler. The MPI-ESM model was used for CMIP5 inter-comparison project and is the core of the MiKlip Prototype Prediction System for climate forecast (Pohlmann et al., 2004; Mueller et al., 2012; Marotzke et al., 2016).

Four ensemble generation methods are used to produce retrospective predictions. Except for the Ensemble Kalman filter based hindcast experiment, the initialisation of the model with observations is the same as in the Baseline 1 hindcasts as described in Mueller et al., 2012 and Pohlmann et al., 2013.

- Anomaly transform (AT) as the first method is based on orthogonalisation of the preliminary prepared anomaly patterns for ten years from the operational ocean re-analysis data ORAS4 (detailed description in Keller et al., 2010; Romanova and Hense, 2015). The temperature, salinity and zonal and meridional velocities are used to calculate the kinetic and potential energy. An energy matrix is constructed such that it contains all the spatial and temporal dimensions of the energy components. From this data matrix the typical pattern modes are extracted by calculating the extended EOFs. The first five modes with positive and negative sign are used to derive the perturbations around the initialized state. They are added to the fields of salinity, temperature and the horizontal velocities. Additional scaling by constraining the patterns in amplitude to the ocean re-analysis GECCO2 (Köhl, 2015) was necessary to derive an appropriate perturbation amplitude (for details see Romanova and Hense, 2015). The hindcasts were performed starting from the ten perturbed ocean initial conditions (five modes with opposite amplitudes) with the atmosphere leaving unperturbed. However only 8 were used for the analysis to match the ensemble size of the Baseline 1 and the EnKF hindcasts.
- Baseline 1 is the hindcast produced from the MiKlip Prediction system (Pohlmann et al., 2004; Mueller et al., 2012). It has been comprehensively studied and evaluated by the MiKlip community in Germany (Kaspar et al., 2016). It uses a simplified ensemble generation scheme, in which the ocean and the atmosphere perturbations are taken from eight sequent days from 1st of January of each year from the assimilation run. However, this scheme is thought to provide too small initial spread. Another problem that might emerge in this case, is that the disturbances of the atmosphere are strongly related on short weather timescales.
- Equivalently to the Baseline 1, the Ocean Lagged hindcast is performed using 10 lagged days from 1st of January. The difference to the Baseline 1 is that the daily altered ocean states were applied to the ocean initial conditions, and the atmospheric perturbations were not taken into account. Such a choice of the initial conditions show very similar spatial structures of the oceanic disturbances, which produce initial spread focused on a stationary ocean state.
- EnKF hindcasts are based on 8 member weakly coupled assimilation with EnKF assimilation of monthly temperature and salinity observations for the ocean (Brune et al., 2015) and the nudging scheme of MiKlip Baseline 1 (Pohlmann et al., 2013) for the atmosphere. For the ocean, we complement the subsurface observations of temperature and salinity from EN4 (Good et al., 2013) with sea surface temperatures from HadISST (Rayner et al., 2003), for the atmospheric nudging we use ERA40 and ERA Interim re-analysis data (Uppala et al., 2005; Dee et al., 2011). The assimilation runs from 1958 to 2014, the hindcast ensemble is initialized every year on January 1st by using the assimilation ensemble last updated on December 31st the year before. In the EnKF initialized hindcast the disturbances are applied simultaneously in atmosphere and ocean. All hindcasts are analysed for different lead years using eight member ensemble.

2.2 Data

In contrast to usual evaluation approaches, where as basic analysis variables near surface temperature or free atmosphere parameters like geopotential height are chosen, we will use the freshwater flux $P-E$ which couples the atmospheric branch of the hydrological cycle to the oceanic one. The aim is to perform an evaluation which is process based. This will enable us to draw conclusions about the hindcasts with respect to the hydrological cycle.

The NCEP R2 atmospheric re-analysis is used for the evaluation of the four different hindcast ensembles. NCEP R2 is an atmospheric re-analysis produced by NOAA National Centers for Environmental Predictions (Kanamitsu et al., 2002). It is the second version of their first NCEP R1 re-analysis (Kalnay et al., 1996) starting from the beginning of the major satellite era. The model is an atmospheric spectral model on a Gaussian grid with resolution T62 (209 km) with 28 vertical sigma levels. The R2 re-analysis release uses observed precipitation forcing which was not included in NCEP R1.

The NCEP R2 re-analysis product covers the same time period starting in the year 2000 until 2011. The global mean for that time period are 0.02 mm/day for NCEP R2. Prior to evaluation of the global mean, climatological seasonal cycle and decadal linear trends were removed. Fig. 1a,b,c shows the grid pointwise estimation of correlation coefficients to the atmospheric re-analysis MERRA (GEOS-5 from NASA/GMAO, Wong et al., 2011), to the ocenaic re-analysis GECCO2 (Köhl, 2015), and to the coupled data assimilation product from GFDL (Zhang et al., 2007; Chang et al., 2013). The two atmospheric reanalyses, NCEP R2 and MERRA, are closest to each other, having an avarage correlation coefficient larger than 0.6. The other two data sets GECCO2

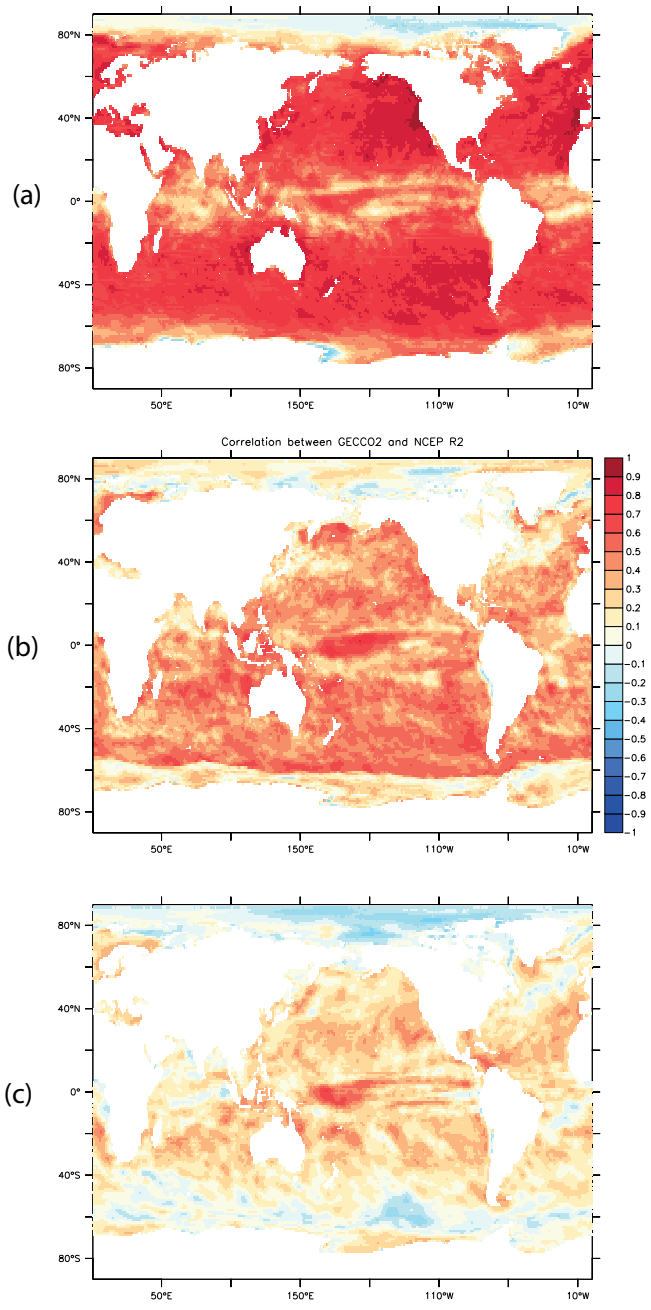

Figure 1: The correlation paterns of NCEP R2 with a) MERRA atmospheric re-analysis, b) GECCO2 ocean re-analysis and c) GFDL coupled re-analysis. Prior, the product mean, the climatological seasonal cycle and the linear trend were removed.

and GFDL clearly depart from this with average correlations relative to NCEP R2 of 0.4 and even less. Also the correlations between GECCO2 and GFDL are low, indicating that these data sets probably contain other source of uncertainty compared to NCEP R2 and MERRA. Additionally NCEP R2 was explicitly produced with the background to provide a better estimate of the energy and water cycle (Kanamitsu et al., 2002). Therefore we will use NCEP R2 as the basic data set estimating the freshwater fluxes in the observed system.

3 Evaluation of the hindcasts

The quality of the different ensemble hindcasts is assessed on unbiased monthly means of the air-sea freshwater flux $P - E$.

3.1 Mean climate state statistics

Standard Deviation

The variability of the four hindcasts, measured through standard deviation around their ensemble means (Fig. 2b, c,d,e) is suppressed in the North and equatorial Atlantic for the different lead times. However, the variability of the freshwater fluxes increases after a forecast lead time of five years for all retrospective forecasts. The largest increases in the order of half of the mean amplitude, are found particularly in the Indian Ocean. With respect to the different ensemble generation schemes, the EnKF filter hindcast exhibit the strongest increases in variability with the hindcast lead times.

Cumulative frequency analysis

The cumulative distribution function (CDF) is the estimated probability that a realization of either the forecast X_{hind} or observation X_{obs} is less equal than a fixed threshold x. If as thresholds the realization values themselves are taken, the CDF $D(x)$ can be simply obtained by ranking the full data set in ascending order.

$$\begin{aligned} D_{\text{hind}}(x) &= \text{Prob}(X_{\text{hind}} \leq x) \\ D_{\text{obs}}(x) &= \text{Prob}(X_{\text{obs}} \leq x) \end{aligned} \tag{3.1}$$

The $D(x)$ functions for the 8 ensemble members of the hindcasts and the NCEP R2 re-analysis data for different lead years are calculated from all grid point values over the verification period after removal of overall mean, the annual cycle and the linear trend. The coherence between the predictions and the observations (shown in Fig. 3 for EnKF hindcast) is different for the different lead years. The CDF from the NCEP R2 stays within the ensemble spread for the investigated time intervals.

To obtain a quantitative measure of the differences and of the quality of the hindcast toward the re-analysis for different lead years the square root of the integrated squared distance Δ_I of the CDF's (Thorarinsdottir et al., 2013) is calculated

$$\Delta_I = \sqrt{\sum_x (D_{\text{hind}}(x) - D_{\text{obs}}(x))^2}. \tag{3.2}$$

Defined in this way, the score shows in one number the departure of the ensemble from the NCEP R2 data (Fig. 4). When Δ_I is close to zero the hindcasts and observations share an identical CDF or are calibrated. However in the present case the results from the score emphasizes more on the quality of the assimilation data set than the quality of the hindcast. The score can not

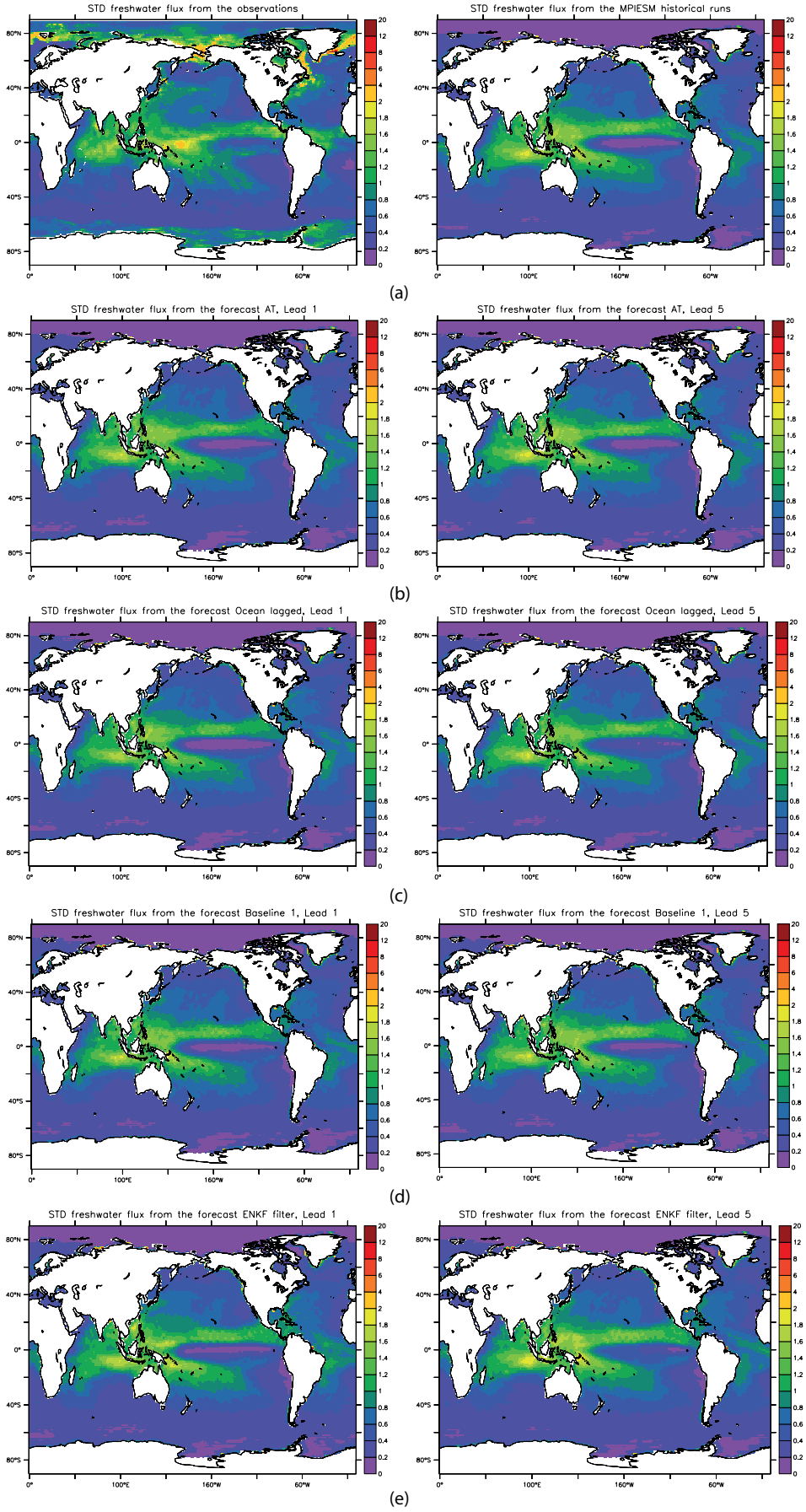

Figure 2: The spatial pattern of the standard deviation around the mean (a) of the four re-analysis products GFDL, NCEP R2, GECCO2 and MERRA and the MPIESM historical runs; and for lead year 1 and lead year 5 (b) AT, (c) Ocean lagged, (d) Baseline 1, and (e) EnKF hindcasts. The unit is mm/day.

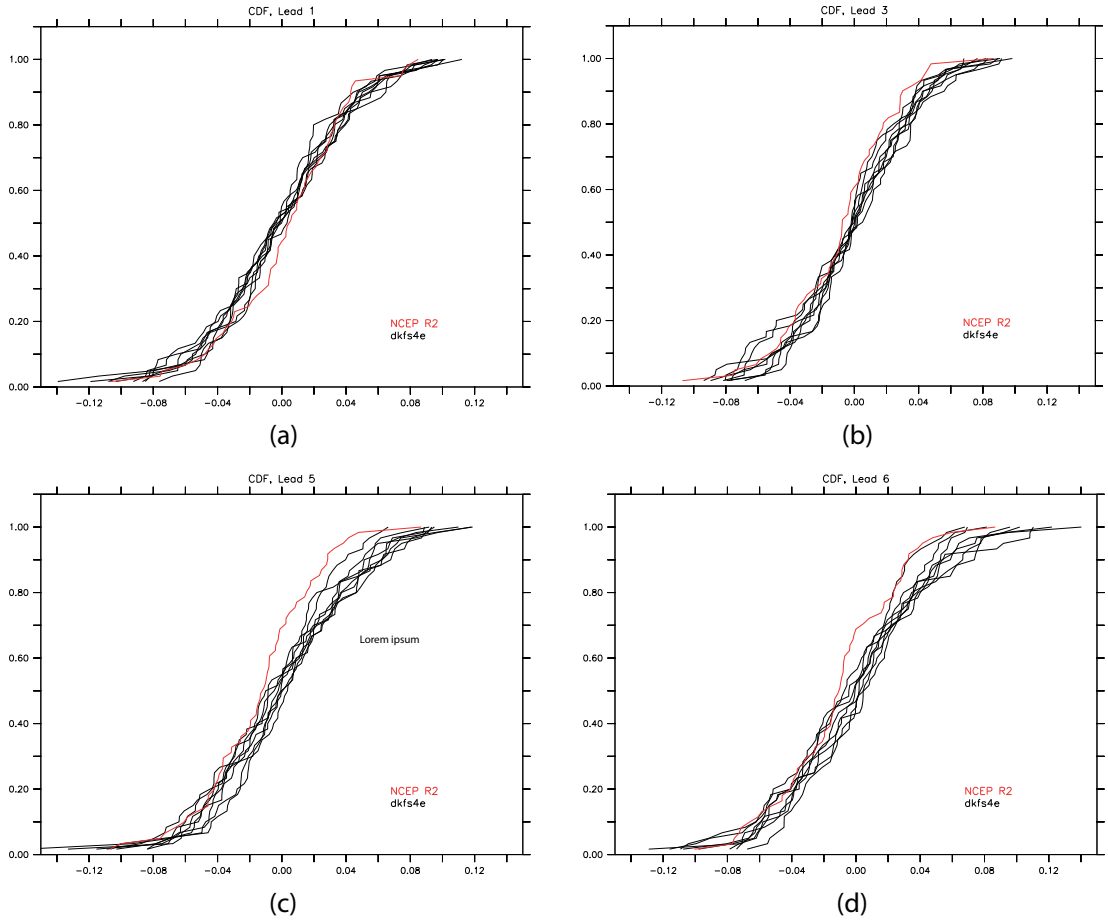

Figure 3: Cumulative frequency distribution function for the 8 members of the EnKF hindcast (probability estimates) for the lead year 1, 3, 5 and 6 (black curves) and the NCEP R2 re-analysis. The lead years and the reanalyses data cover the same time window.

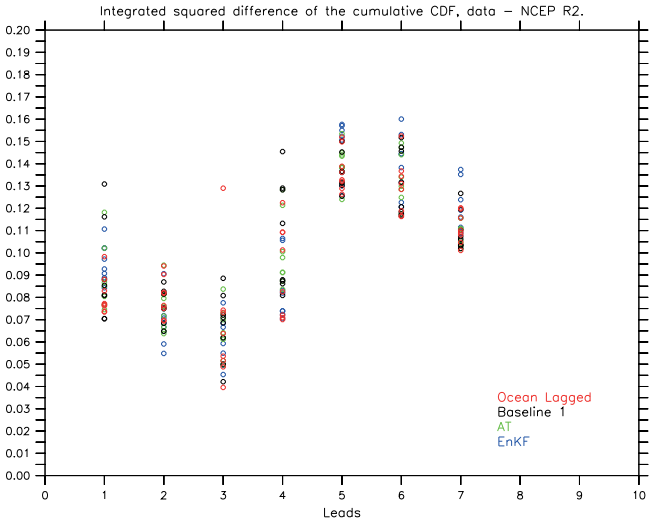

Figure 4: Integrated squared difference of each ensemble member of the EnKF, Baseline 1, Ocean lagged and AT hindcasts cumulative frequency distribution and the NCEP R2 re-analysis data.

give clear sign for a predominance of a certain ensemble generation scheme. However, it shows at which lead time the prediction fits at best the re-analysis data set. This score is more useful to estimate the forecasted deviations from different data sets.

Table 1: Percentage of grid points with correlation to NCEP R2 greater than 0.3.

	Lead year 1	Lead year 5	Lead year 2 to 5
AT	1.81	2.58	**5.40**
Lagged Ocean	2.13	**2.63**	4.75
Baseline 1	2.37	2.33	4.97
EnKF	**4.51**	2.50	3.54

3.2 Probabilistic evaluation

Correlation analysis

The spatial distribution of the correlation coefficients significant at the 5 % level between the monthly averages of the ensemble means of the hindcasts and the NCEP R2 are shown in Fig. 5 and Fig. 6. The percentages of grid points with coefficients larger than 0.3 are listed in Table 1.

EnKF hindcast shows the greatest area, approximately two times larger compared to the other hindcasts, with coefficient exceeding 0.3 for the lead year 1 and for all data sets. For the lead year 5, the Ocean Lagged hindcast results in the largest area of significant correlation, with coefficients over 0.3. A decrease of the correlation with the lead years is found for the EnKF hindcast. This

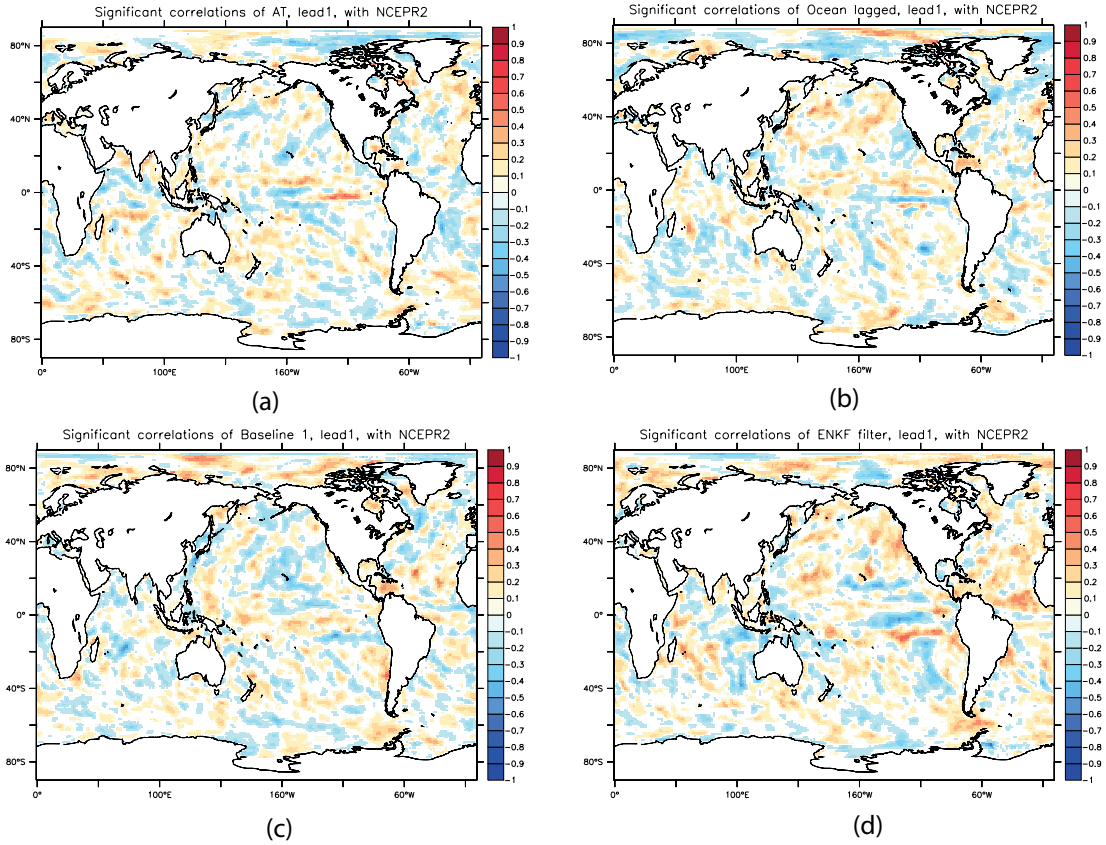

Figure 5: Spatial pattern of the correlation coefficients between NCEP R2 and ensemble mean of different hindcasts for the lead year 1 at 5 % significance level.

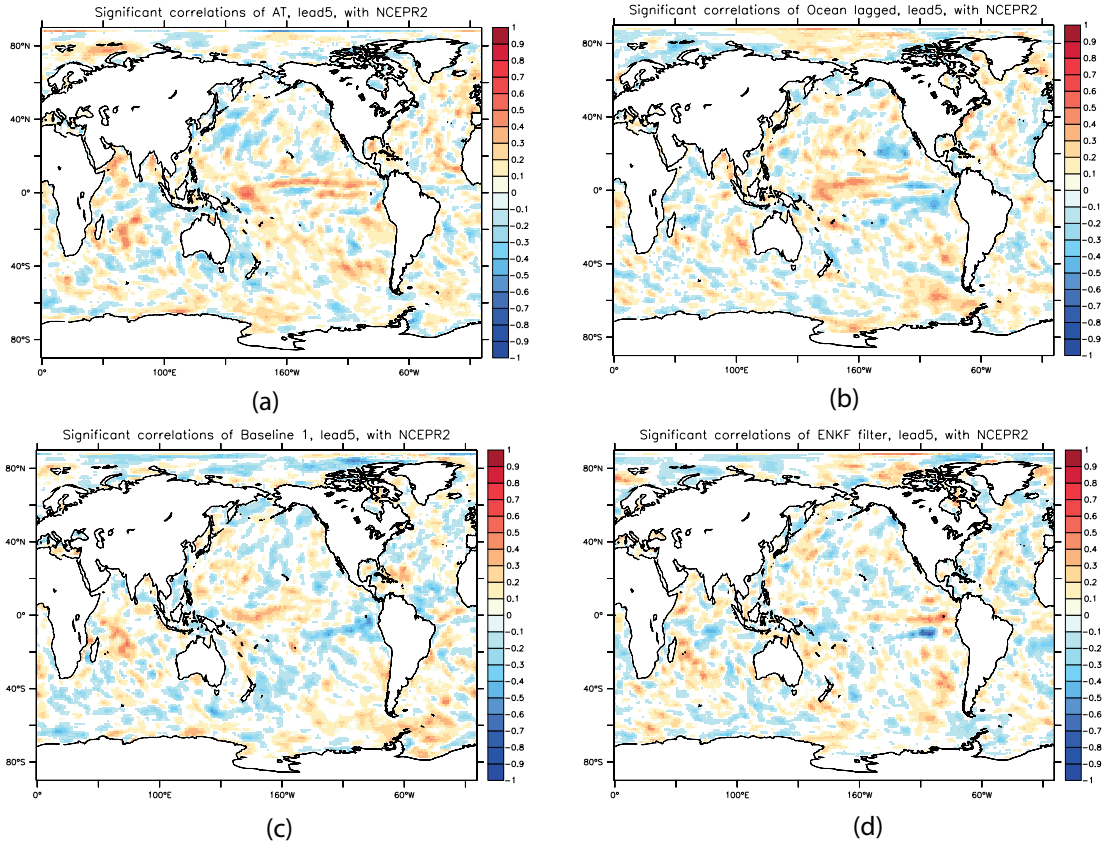

Figure 6: The same as in Fig. 5 but for the lead year 5.

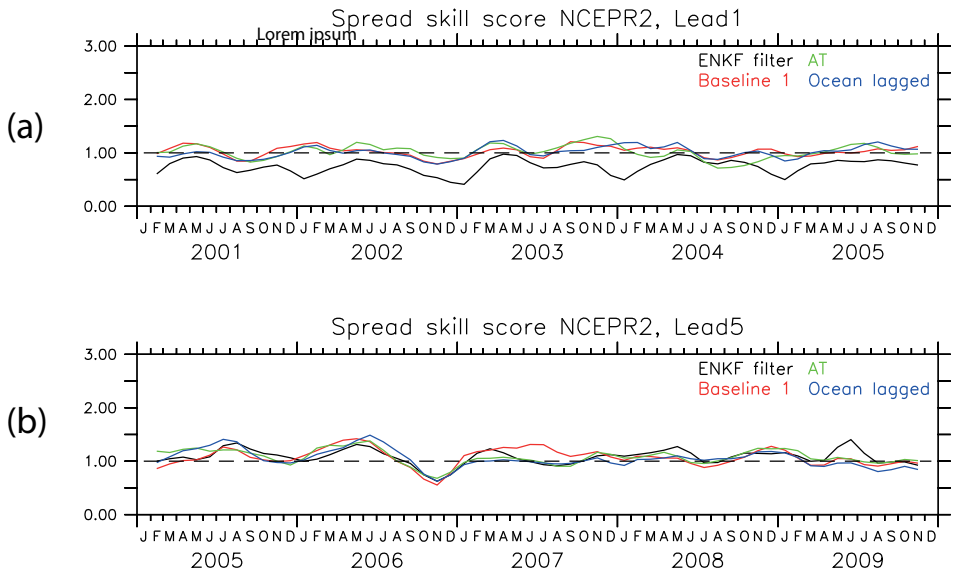

Figure 7: Ensemble spread score calculated for the air-sea freshwater fluxes produced from EnKF, Baseline 1, Ocean lagged, and AT hindcast for the lead 1 and 5. The NCEP R2 re-analysis is considered as the truth.

behavior of the EnKF hindcast is related to the strong variability increase with the lead years which has been shown in Fig. 2. However, for AT hindcast, an increase of the correlation on the inter-annual to multi-annual time-scales is found when going from lead year 1 to lead year 5. The overall significant correlation increases when averaging the lead years 2 to 5. The highest values are in favour for the AT ensemble generation method (Table 1).

Spread skill score

Another metric to assess the quality of the hindcast relates the variability within the ensemble to the mean square error. This tests if the observations are similar to a single realisation of the ensemble relative to the ensemble mean. The ratio of the global means of within-ensemble variability to the mean square error between the ensemble mean and the observation is called the ensemble spread score (Keller et al., 2010).

$$E_{\text{SpreadSkill}} = \frac{\sqrt{\sum_{t,k}(X_{t,k} - X_{t,\text{ens}_{\text{mean}}})^2}}{\sqrt{\sum_{t}(X_{t,obs} - X_{t,\text{ens}_{\text{mean}}})^2}} \qquad (3.3)$$

with the sum taken over verification time interval (index t) and the number of ensemble realisations k. The score takes a value of one, when the ensemble spread is perfect. If the $E_{\text{SpreadSkill}}$ is less (larger) than 1 the spread is under-(over)estimated.

The results of the ESS for lead years 1 and 5 are shown in Fig. 7. Evident improvement of the score is seen only for EnKF for the lead year 5. But similarly as the CDF analysis, this score actually shows no large difference between the hindcasts. Although EnKF hindcasts show a behavior different than the other hindcast, this still does not point to some advantage or disadvantage of the selected method.

Reliability diagrams

Measuring the reliability or calibration of a forecast means that cases with a specified predicted probability of occurrence of an event occur with that frequency in the corresponding observations: cases where the occurrence of an event is predicted to be unlikely should also occur only rarely in the verifying observations and vice verse. Here we consider as event the occurrence of a positive freshwater flux anomaly.

A graphical presentation is the reliability diagram which maps the observed frequency of the event during situation when a certain probability of the event P_f is predicted. A reliable forecast system is given if the observed frequencies align along the diagonal meaning that on average over the full data set the predicted probabilities P_f are identical to the observed frequencies conditioned on the predicted probabilities.

Reliability of a forecast should be based on a large sample size, which the hindcasts can not provide for a single model grid point or area mean with 5 years 12 months data length. Instead geographical boxes of 10° by 10° are considered which provides up to 60 000 events (or less, when the box includes land points) in the data series.

Fig. 8 shows an example of the reliability diagrams for the Tropical North Atlantic for all the hindcast and for lead year 1 and 5. A histogram of predicted probabilities P_f is also included in the plot. If the curve of the observed relative frequency lies below the diagonal line, this indicates overforecasting (probabilities too high), if the curve lies above the line indicate underforecasting (probabilities too low). When the curve is exactly along the diagonal line, the forecast system is called reliable. Most of the diagrams on 10° by 10° sliding box exhibit underforecasting (not shown). To obtain a single number characterizing the reliability a linear regression of the

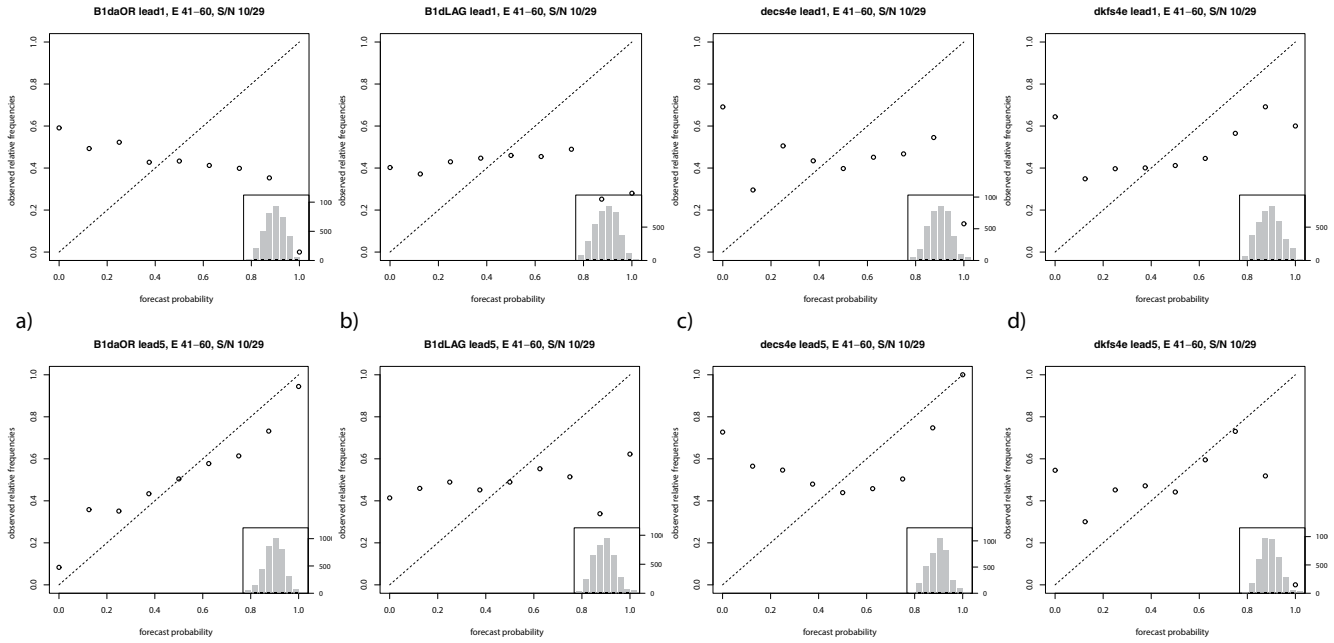

Figure 8: Example reliability diagrams for a 10° 10° box over the Tropical North Atlantic for the leads 1 (upper panel) and 5 (lower panel) for a) AT, b) Lagged Ocean, c) Baseline 1 and d) EnKF.

frequency curve is constructed and the angle between the regression line and the diagonal is calculated with:

$$\tan\theta = \frac{l_1 - l_2}{1 + l_1 * l_2}, \tag{3.4}$$

where $l_1 = 1$ and l_2 are the slopes of the diagonal and the linear fit. If θ is in the interval $[-30°, 30°]$ we consider the forecast reliable. If θ is lower than $-60°$ or larger than 60° the forecast is unreliable, when θ is between $-30°$ and $-60°$ or between 30° and 60° reliability is not well defined. The reliability for all the hindcast and observations is calculated within a moving box of 10° Lat by 10° Lon and θ is assigned to the central point of that window and plotted at each point for the lead years 1 and 5 (Fig. 9 and Fig. 10). The reliability of the hindcast depends on the given verification data set and reliable forecasts are shaded in green, yellow shading indicates the not well defined cases, while red is an indication of unreliability (Weisheimer and Palmer, 2014; Stolzenberger et al., 2015).

Fig. 9 and 10 do not show any obvious spatially coherent signal for reliability even more than the correlation patterns in Fig. 5 and 6. This could be the result of the rather large uncertainties in the re-analysis freshwater fluxes or of missing quality in the hindcasts. To measure reliability on a global scale we evaluated the global percentage of reliable (green areas) versus unreliable forecasts. These numbers are shown in Table 2 and Table 3. At this global level of aggregation the evaluation of the hindcasts at the global scale can be summarized as follows: the globally aggregated reliability related skill for the lead year 1 reaches 21 % for the Ocean Lagged hindcast. If one compares the hindcasts

Table 2: Percentage reliable forecast for different ensemble generation experiments compared to different NCEP R2 re-analysis data set.

	Lead year 1	Lead year 5
AT	18.31	**25.00**
Lagged Ocean	**21.17**	19.13
Baseline 1	14.52	14.97
EnKF	17.06	15.43
Baseline 1	14.52	14.97

Table 3: Percentage un-reliable forecast for different ensemble generation experiments compared to NCEP R2 re-analysis data set.

	Lead year 1	Lead year 5
AT	12.32	**14.13**
Lagged Ocean	12.65	15.39
Baseline 1	15.81	21.18
EnKF	**6.13**	15.32

produced with different ensemble generation schemes, the AT ensemble generation method exhibits the largest values (bold numbers in Table 2) for the lead year 5 to 25 %. Considering on the global scale the percentages of the unreliable forecast (red areas in Fig. 9 and 10) show again a consistent picture. The smallest values are obtained from the hindcasts which use the EnKF filter for defining the initial conditions, (Table 3). This means that the two physically based ensemble generation methods provide a small but persistent signal on the global scale.

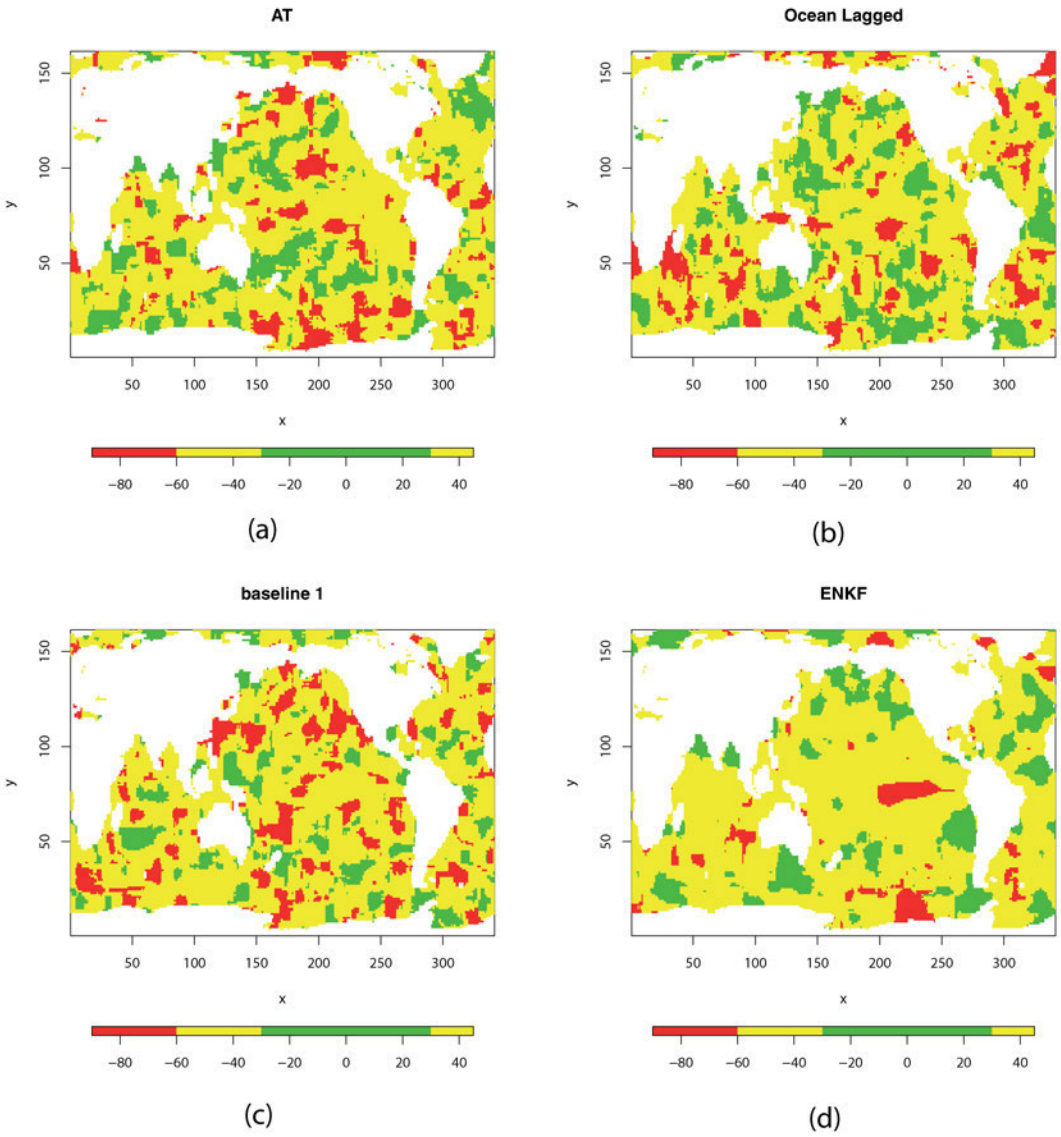

Figure 9: Horizontal assessment of reliability of the four hindcasts, lead 1, towards NCEP R2 re-analysis data. The green color shows reliable diagram e.g. the frequency observation probability versus probability estimates lies in an angle of −30° to 30° with respect to perfect prediction. The yellow color is undefined forecast and the red is unreliable forecast.

4 Discussion and conclusion

To test and assess the skill of the ensemble generation method we performed two hindcast experiments and investigated another two ones. Each of the experiments uses different schemes for creating the initial perturbation patterns. The ensemble generation methods differ in two main characteristics. One feature is that:

- two of the hindcasts (AT and Ocean Lagged) disturb initial conditions only in the ocean, since the interannual and decadal variability are a product of the long term ocean dynamics
- the other two hindcasts perturb the atmosphere and ocean simultaneously, accounting for synchronized coupled error definitions (Baseline 1 and EnKF).

The second feature is based on the annotation of the phase space of the perturbations at the initial state of the ensemble runs. One of the hindcast is based on initial perturbations using orthogonal anomaly fields of the flow and the density fields while the other is based on patterns spanning the multivariate probability density of the initial state. The first are derived from physically consistent anomalies generated by the internal ocean dynamics (AT) while the second are determined by the observations and the dynamics of the ocean model through the EnKF approach. Both force each single realisation in quite different ways than the lagged initial disturbances resulting from the very different pattern structure. The variable which is tested in this study is the air-sea freshwater fluxes which knowledge is an important parameter for the coupling of atmosphere and ocean. With respect to the significance of the air-sea flux for

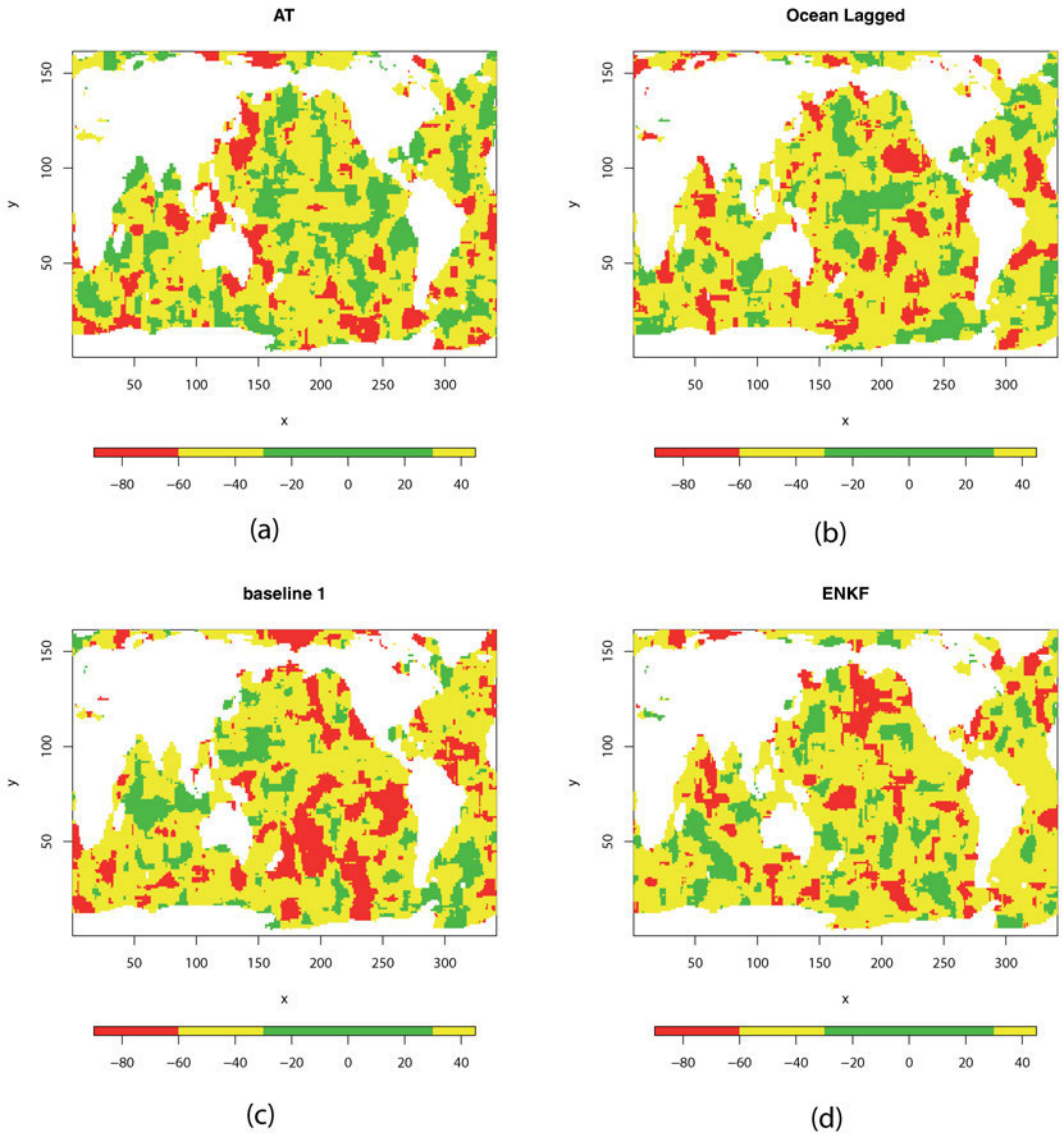

Figure 10: The same as in Fig. 9 but for the lead year 5.

climate prediction, the task appears to be challenging since the re-analysis data is uncertain on the exact geographical location of the occurrence of the variability and its strength. In this study, we investigate the behavior of the hindcasts using different evaluation scores. Although not all of the scores can point to explicit dominance of the skill of one or another ensemble generation method, we search for the most sensible ones, since the differences between the ensemble generation methods are small. The validation of the hindcasts considers the interannual and multi-year variability and does not account for the decadal linear trends and the seasonal cycle. Validation of hindcasts on time scales larger than years has currently been done without considering uncertainties in the verifying re-analysis. The climate prediction of freshwater fluxes, which collect the uncertainties from the model limitations (parametrization schemes) combined with the quasi observed freshwater data sets, is not yet in a stage to provide answers about decadal climate predictabitlity. In summary here, we can list our findings as follows:

- Comparing the skill of the investigated hindcast with respect to the disturbed system component, ocean alone or ocean and atmosphere, we do not find any significant effect when the atmosphere is perturbed. This holds for the regional analysis as well as for the global aggregations. It does not mean that the initialization of the atmosphere with observations is unimportant but that the uncertainty associated with the atmospheric initial fields could probably be ignored. The increase of the error growth rate with the lead years is more due to the ocean dynamics than due to the inherent atmospheric instabilities which generate the uncertainties.
- The most successful forecast in terms of percentage of positive correlations larger than 0.3 between re-analysis and hindcasts and spatial reliable esti-

mates and smaller percentages of unreliable predictions is in favour for the advanced ensemble generation methods Anomaly Transform and Ensemble Kalman filtering. AT creates the largest error spread at the beginning of the ensemble runs while EnKF provides initial disturbances which are defined by the error structure of the observations and the internal uncertainty of the dynamic model. Except for the spatial disturbance orientation, an important role also may play the norm on which the perturbation patterns are created and the rescaling method for the amplitudes (Romanova and Hense, 2015). But more important both methods incorporate dynamical information of uncertainty structures which might give a hint on their slight advantages. The improvement of both AT and EnKF versus the lagged initializations (Ocean Lagged and Baseline 1) is found to be a feature when comparing to the NCEP R2 re-analysis of the freshwater fluxes. But despite this advantage the overall gain in skill is still small of about 5 %.

In conclusion, different verification methods not always give an explicit and one side answer to the questions concerning the skillfulness of the forecast or the advantage of different ensemble generation methods since either possess sensitivity in one or another aspect. Other difficulties arise from the large uncertainties in the observational data set. Our study does not show a huge profit of one or another ensemble generation scheme on interannual timescales. Similar results have been found when comparing different ensemble generation methods with other types of data like sea surface temperature. But here the analysis often stopped at regional scales. Having done an aggregation of skill measures to larger scales e.g. global numbers for the freshwater fluxes, we found that the above results give hints for the direction of further research development, either in expanding the initial error space phase based on physical modes or in improving the evaluation methods by an explicit treatment of different scales.

Acknowledgments

This work was funded by the German Federal Ministry of Research and Higher Education (BMBF) within the MiKlip program Module A through the AODA-PENG project (FKZ 01LP1157A). The calculations were performed at the German Climate Computing Center (DKRZ) in Hamburg and for the evaluation purposes the Standard VECAP tool and SPECS evaluation tools were used.

References

Adler, R.F., C. Kidd, G. Petty, M. Morissey, M. Goodman, 2001: Intercomparison of global precipitation products: The third precipitation intercomparison project (PIP-3). – Bull. Amer. Meteor. Soc. **82**, 1377–1396, DOI: 10.1175/1520-0477(2001)082<1377:IOGPPT>2.3.CO.

Béranger, K., L. Siefridt, B. Barnier, E. Garnier, H. Roquet, 1999: Evaluation of operational ECMWF surface freshwater fluxes over oceans during 1991–1997. – J. Marine Sys. **22**, 13–36, DOI: 10.1016/S0924-7963(99)00028-7.

Broecker, W.S., 1994: Massive iceberg discharges as triggers for global climate change. – Nature **372**, 421–424.

Brune, S., L. Nerger, J. Baehr, 2015: Assimilation of oceanic observations in a global coupled earth system model with the {SEIK} filter. – Ocean Modelling **96**, Part 2, 254–264, DOI: 10.1016/j.ocemod.2015.09.011.

Chang, Y.-S., S. Zhang, A. Rosati, T. Delworth, W. Stern, 2013: An assessment of oceanic variability for 1960–2010 from the GFDL ensemble coupled data assimilation. – Climate Dyn. **40**, 775–803, DOI: 10.1007/s00382-012-1412-2.

Dee, D.P., S.M. Uppala, A.J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M.A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A.C.M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A.J. Geer, L. Haimberger, S.B. Healy, H. Hersbach, E.V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A.P. McNally, B.M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, F. Vitart, 2011: The era-interim reanalysis: configuration and performance of the data assimilation system. – Quart. J. Roy. Meteor. Soc. **137**, 553–597, DOI: 10.1002/qj.828.

Du, H., F. Doblas-Reyes, J. García-Serrano, V. Guemas, Y. Soufflet, B. Wouters, 2012: Sensitivity of decadal predictions to the initial atmospheric and oceanic perturbations. – Climate Dyn. **39**, 2013–2023, DOI: 10.1007/s00382-011-1285-9.

Giglio, D., D. Roemmich, 2014: Climatological monthly heat and freshwater flux estimates on a global scale from Argo. – J. Geophys. Res. Oceans **119**, 6884–6899, DOI: 10.1002/2014JC010083.

Giorgetta, M.A., J. Jungclaus, C.H. Reick, S. Legutke, J. Bader, M. Böttinger, V. Brovkin, T. Crueger, M. Esch, K. Fieg, K. Glushak, V. Gayler, H. Haak, H.-D. Hollweg, T. Ilyina, S. Kinne, L. Kornblueh, D. Matei, T. Mauritsen, U. Mikolajewicz, W. Müller, D. Notz, F. Pithan, T. Raddatz, S. Rast, R. Redler, E. Roeckner, H. Schmidt, R. Schnur, J. Segschneider, K.D. Six, M. Stockhause, C. Timmreck, J. Wegner, H. Widmann, K.-H. Wieners, M. Claussen, J. Marotzke, B. Stevens, 2013: Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5. – J. Adv. Mod. Earth Sys. **5**, 572–597, DOI: 10.1002/jame.20038.

Good, S.A., M.J. Martin, N.A. Rayner, 2013: En4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. – J. Geophys. Res. Oceans **118**, 6704–6716, DOI: 10.1002/2013JC009067.

Haughton, N., G. Abramowitz, A. Pitman, S. Phipps, 2014: On the generation of climate model ensembles. – Climate Dyn. **43**, 2297–2308, DOI: 10.1007/s00382-014-2054-3.

Iwasaki, S., M. Kubota, T. Watabe, 2014: Assessment of various global freshwater flux products for the global ice-free oceans. – Remote Sens. Env. **140**, 549–561, DOI: 10.1016/j.rse.2013.09.026.

Josey, S.A., R. Marsh, 2005: Surface freshwater flux variability and recent freshening of the North Atlantic in the eastern subpolar gyre. – J. Geophys. Res. Oceans **110**, DOI: 10.1029/2004JC002521.

Jungclaus, J.H., N. Fischer, H. Haak, K. Lohmann, J. Marotzke, D. Matei, U. Mikolajewicz, D. Notz, J.-S. von Storch, 2013: Characteristics of the ocean simulations in MPIOM, the ocean component of the MPI earth

system model. – J. Adv. Model. Earth Sys. **5**, 422–446, DOI: 10.1002/jame.20023.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K.C. Mo, C. Ropelewski, J. Wang, R. Jenne, D. Joseph, 1996: The NCEP/NCAR 40-year reanalysis project. – Bull. Amer. Meteor. Soc. **77**, 437–471, DOI: 10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Kanamitsu, M., W. Ebisuzaki, J. Woollen, S.-K. Yang, J.J. Hnilo, M. Fiorino, G.L. Potter, 2002: NCEP-DOE AMIP-II reanalysis (R-2). – Bull. Amer. Meteor. Soc. **83**, 1631–1643, DOI: 10.1175/BAMS-83-11-1631.

Kaspar, F., H.W. Rust, U. Ulbrich, P. Becker, 2016: Verification and process oriented validation of the MIKLIP decadal prediction system. – Meteorol. Z. **25**, 629–630, DOI: 10.1127/metz/2016/0831.

Keenlyside, N., M. Latif, J. Jungclaus, L. Kornblueh, E. Roeckner, 2008: Advancing decadal-scale climate prediction in the North Atlantic sector. – Nature **453**, 84–88.

Keller, J.D., A. Hense, L. Kornblueh, A. Rhodin, 2010: On the orthogonalization of bred vectors. – Wea. Forecast. **25**, 1219–1234.

Köhl, A., 2015: Evaluation of the GECCO2 ocean synthesis: transports of volume, heat and freshwater in the atlantic. – Quart. J. Roy. Meteor. Soc. **141**, 166–181, DOI: 10.1002/qj.2347.

Kruschke, T., H.W. Rust, C. Kadow, W.A. Müller, H. Pohlmann, G.C. Leckebusch, U. Ulbrich, 2015: Probabilistic evaluation of decadal prediction skill regarding northern hemisphere winter storms. – Meteorol. Z. **25**, 721–738, DOI: 10.1127/metz/2015/0641.

Magnusson, L., M. Leutbecher, E. Källén, 2008: Comparison between singular vectors and breeding vectors as initial perturbations for the ECMWF ensemble prediction system. – Mon. Wea. Rev. **136**, 4092–4104, DOI: 10.1175/2008MWR2498.1.

Marini, C., I. Polkova, A. Köhl, D. Stammer, 2016: A comparison of two ensemble generation methods using oceanic singular vectors and atmospheric lagged initialization for decadal climate prediction. – Mon. Wea. Rev. **144**, 2719–2738.

Marotzke, J., W. Müller, F.S.E. Vamborg, P. Becker, U. Cubasch, H. Feldmann, F. Kaspar, C. Kottmeier, C. Marini, I. Polkova, K. Prömmel, H. Rust, D. Stammer, U. Ulbrich, C. Kadow, A. Köhl, J. Kröger, T. Kruschke, J. Pinto, H. Pohlmann, M. Reyers, M. Schröder, F. Sienz, C. Timmreck, M. Ziese, 2016: MIKLIP – a national research project on decadal climate prediction. – Bull. Amer. Meteor. Soc., published online, DOI: 10.1175/BAMS-D-15-00184.1.

Matei, D., J. Baehr, J. Jungclaus, H. Haak, W. Müller, J. Marotzke, 2012: Multiyear prediction of monthly mean atlantic meridional overturning circulation at 26.5 N. – Science **335**, 76–79.

Molteni, F., R. Buizza, T.N. Palmer, T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. – Quart. J. Roy. Meteor. Soc. **122**, 73–119, DOI: 10.1002/qj.49712252905.

Mueller, W.A., J. Baehr, H. Haak, J.H. Jungclaus, J. Kröger, D. Matei, D. Notz, H. Pohlmann, J. von Storch, J.S. Marotzke, 2012: Forecast skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology. – Geophys. Res. Lett. **39**, L22707 DOI: 10.1029/2012GL053326.

Pohlmann, H., M. Botzet, M. Latif, A. Roesch, M. Wild, P. Tschuck, 2004: Estimating the decadal predictability of a coupled aogcm. – J. Climate **17**, 4463–4472.

Pohlmann, H., W.A. Müller, K. Kulkarni, M. Kameswarrao, D. Matei, F.S.E. Vamborg, C. Kadow, S. Illing, J. Marotzke, 2013: Improved forecast skill in the tropics in the new MIKLIP decadal climate predictions. – Geophys. Res. Lett. **40**, 5798–5802, DOI: 10.1002/2013GL058051.

Rayner, N.A., D.E. Parker, E.B. Horton, C.K. Folland, L.V. Alexander, D.P. Rowell, E.C. Kent, A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. – J. Geophys. Res. Atmos. **108**, 4407, DOI: 10.1029/2002JD002670.

Romanova, V., A. Hense, 2015: Anomaly transform methods based on total energy and ocean heat content norms for generating ocean dynamic disturbances for ensemble climate forecasts. – Climate Dyn. **48**, 1–24, DOI: 10.1007/s00382-015-2567-4.

Romanova, V., A. Köhl, D. Stammer, C. Klepp, A. Andersson, S. Bakan, 2010: Sea surface freshwater flux estimates from GECCO, HOAPS and NCEP. – Tellus A **62**, 435–452, DOI: 10.1111/j.1600-0870.2010.00447.x.

Smith, D., S. Cusack, A. Colman, C. Folland, G. Harris, J. Murphy, 2007: Improved surface temperature prediction for the coming decade from a global climate model. – Science **317**, 796–799.

Stammer, D., K. Ueyoshi, A. Köhl, W.G. Large, S.A. Josey, C. Wunsch, 2004: Estimating air-sea fluxes of heat, freshwater, and momentum through global ocean data assimilation. – J. Geophys. Res. Oceans **109**, C05023, DOI: 10.1029/2003JC002082.

Stevens, B., M. Giorgetta, M. Esch, T. Mauritsen, T. Crueger, S. Rast, M. Salzmann, H. Schmidt, J. Bader, K. Block, R. Brokopf, I. Fast, S. Kinne, L. Kornblueh, U. Lohmann, R. Pincus, T. Reichler, E. Roeckner, 2013: Atmospheric component of the MPI-M Earth System Model: ECHAM6. – J. Adv. Mod. Earth Sys. **5**, 146–172, DOI: 10.1002/jame.20015.

Stolzenberger, S., R. Glowienka-Hense, T. Spangehl, M. Schröder, A. Mazurkiewicz, A. Hense, 2015: Revealing skill of the MiKlip decadal prediction system by three-dimensional probabilistic evaluation. – Meteorol. Z. **25**, 657–671, DOI: 10.1127/metz/2015/0606.

Thorarinsdottir, T.L., T. Gneiting, N. Gissibl, 2013: Using proper divergence functions to evaluate climate models. – J. Uncertainty Quantification **1**, 522–534.

Toth, Z., E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. – Bull. Amer. Meteor. Soc. **74**, 2317–2330.

Trenberth, K.E., C.J. Guillemot, 1998: Evaluation of the atmospheric moisture and hydrological cycle in the NCEP/NCAR reanalyses. – Climate Dyn. **14**, 213–231.

Uppala, S.M., P.W. Kallberg, A.J. Simmons, U. Andrae, V.D.C. Bechtold, M. Fiorino, J.K. Gibson, J. Haseler, A. Hernandez, G.A. Kelly, X. Li, K. Onogi, S. Saarinen, N. Sokka, R.P. Allan, E. Andersson, K. Arpe, M.A. Balmaseda, A.C.M. Beljaars, L.V.D. Berg, J. Bidlot, N. Bormann, S. Caires, F. Chevallier, A. Dethof, M. Dragosavac, M. Fisher, M. Fuentes, S. Hagemann, E. Hólm, B.J. Hoskins, L. Isaksen, P.A.E.M. Janssen, R. Jenne, A.P. Mcnally, J.-F. Mahfouf, J.-J. Morcrette, N.A. Rayner, R.W. Saunders, P. Simon, A. Sterl, K.E. Trenberth, A. Untch, D. Vasiljevic, P. Viterbo, J. Woollen, 2005: The ERA-40 re-analysis. – Quart. J. Roy. Meteor. Soc. **131**, 2961–3012, DOI: 10.1256/qj.04.176.

Vinogradova, N.T., R.M. Ponte, 2013: Clarifying the link between surface salinity and freshwater fluxes on monthly to interannual time scales. – J. Geophys. Res. Oceans **118**, 3190–3201, DOI: 10.1002/jgrc.20200.

Wei, M., Z. Toth, R. Wobus, Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. – Tellus A **60**, 62–79, DOI: 10.1111/j.1600-0870.2007.00273.x.

Weisheimer, A., T.N. Palmer, 2014: On the reliability of seasonal climate forecasts. – J. Roy. Soc. Interface **11**, published online, DOI: 10.1098/rsif.2013.1162.

Wong, S., B.H. Kahn, B. Tian, B.H. Lambrigtsen, H. Ye, 2011: Closing the Global Water Vapor Budget with AIRS Water Vapor, MERRA Reanalysis, TRMM and GPCP Precipitation, and GSSTF Surface Evaporation. – J. Climate **24**, 6307–6321, DOI: 10.1175/2011JCLI4154.1.

Zaucker, F., T.F. Stocker, W.S. Broecker, 1994: Atmospheric freshwater fluxes and their effect on the global thermohaline circulation. – J. Geophys. Res. **99**, 12443–12457.

Zhang, S., M.J. Harrison, A. Rosati, A. Wittenberg, 2007: System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies. – Mon. Wea. Rev. **135**, 3541–3564, DOI: 10.1175/MWR3466.1.