Tellus
SERIES A
DYANAMIC
METEOROLOGY
AND OCEANOGRAPHY
PUBLISHED BY THE INTERNATIONAL METEOROLOGICAL INSTITUTE IN STOCKHOLM

Generating atmospheric forcing perturbations for an ocean data assimilation ensemble

By ISABELLE MIROUZE*, AND ANDREA STORTO, *Centro Euro-Mediterraneo sui Cambiamenti Climatici, Bologna, Italy*

(Manuscript received 23 December 2018; in final form 22 May 2019)

ABSTRACT

Running ensemble of reanalyses or forecasts has proved successful at improving their performances, despite the cost. Generating ensemble simulations requires generating perturbations within the models, and for the assimilated observations and subsidiary conditions. This paper proposes a statistical model to generate atmospheric forcing random perturbations in a flexible and cheap way, for all the variables required to calculate bulk formulae. The training data are a ten-year set of differences between ERA-INTERIM (from ECMWF) and MERRA (from NASA) atmospheric forcing fields, unbiased with a high-pass filter. The model is designed to generate global spatially-varying multi-variate and unbiased perturbations with consistent spatial structures. Based on linear regressions, the model allows for regression coefficients and residual standard deviations to vary with the time of the year. Once defined, the model does not rely on any other external data to generate the perturbations, and can hence be used on- or off-line in the goal of seeding the ensemble more appropriately. Once designed, the model has been validated by comparing three years of generated perturbations to three years of differences between the reanalyses out of the training period. Statistical tests show that the distributions of the sets comply reasonably well, except for precipitation and snow. The major issues come from regions with high kurtosis or where no clear patterns can be established. In terms of structure, the perturbation standard deviation of both sets show similar patterns for all variables.

Keywords: linear regression, statistical model

1. Introduction

Running an ensemble of reanalyses or forecasts has become prevalent nowadays for its ability to provide flow-dependent information on the background-error structure (Lorenc, 2003; Buehner, 2005; Clayton et al., 2013; Oddo et al., 2016). Firstly introduced by Evensen (1994) for sequential data assimilation (Ensemble Kalman Filter; EnKF), ensemble simulations have also been adapted for variational data assimilation in the atmosphere (Lorenc, 2003; Buehner, 2005) as well as in the ocean (Penny et al., 2015; Storto et al., 2018). Long range prediction systems (seasonal to decadal) rely on ensemble forecasts as well to provide a probabilistic picture of long range forecasts.

The key idea of ensemble reanalyses or forecasts is to provide several simulation realisations whose statistics can realistically represent the error associated with the simulation. An important issue is thus to generate perturbations for these different realisations such that the diagnosed ensemble error covariances are representative of the true forecast error covariances. As stated by Turner et al. (2008), the ensemble variability can be obtained by perturbing (i) initial conditions, (ii) model equations and/or (iii) forcing data.

Ensemble generation has been widely studied and has led to different strategies for defining perturbations to add to the initial guess (e.g. Toth and Kalnay, 1997; Evensen, 2003; Zupanski et al., 2006). Perturbing the model equations is a complex problem that generally requires to turn a deterministic model into a probabilistic model. An example is given by Brankart et al. (2015) on the NEMO ocean model. In a data assimilation ensemble system, forcing data include both the observations fed to the assimilation scheme and the subsidiary conditions used for the model, such as boundary condition forcing fields and static ancillary inputs. Although perturbing the observations is not required for an Ensemble Square Root Filter (Whitaker and Hamill, 2002), it is a crucial element for an Ensemble Kalman Filter or an Ensemble 3D- or 4D-VAR. When required, observation

*Corresponding author. e-mail: Isabelle.Mirouze@cerfacs.fr

Citation: Tellus A: 2019, **71**, 1624459, https://doi.org/10.1080/16000870.2019.1624459

perturbations are generally generated from a normal distribution with a variance corresponding to the variance specified for the observation error in the data assimilation scheme. The subsidiary conditions to perturb include the lateral boundary conditions in the case of limited-area models (e.g. Storto and Randriamampianina, 2010) and the atmospheric forcing fields, which is the core of this paper.

In ocean reanalysis or forecast systems, the model needs to be forced by a set of atmospheric fields. These fields can be provided by either atmospheric reanalyses such as the ECMWF ERA-INTERIM reanalysis (Dee et al., 2011), the NCEP/NCAR reanalysis (Kalnay et al., 1996), or Numerical Weather Prediction systems as, for example, in the Met Office operational system FOAM (Forecasting Ocean Assimilation Model; Blockley et al., 2014). Whatever the way they are provided, the atmospheric forcing fields include an error that will propagate through the ocean model. Generating an ensemble of reanalyses or forecasts with perturbations whose distribution represent this error allows its propagated effect to be realistically accounted for.

Estimating and modelling the distribution of the atmospheric forcing error is not an easy task however. The simplest method consists in drawing a random error from a normal distribution with zero mean and a prescribed standard deviation (e.g. Brusdal et al., 2003). In the general framework they developed, Turner et al. (2008) use statistics on observations collected at different weather stations to determine the standard deviation of the perturbations for their ensemble simulations of Port Phillip Bay. Alves and Robert (2005) developed a technique for generating wind stress perturbations for the Tropical Pacific Ocean by calculating the first 50 EOFs (Empirical Orthogonal Functions) of a two-year intra-seasonally filtered differences set between two wind stress products, and by then combining randomly these 50 EOFs. Wind stress perturbations have also been generated by using monthly mean anomalies calculated from differences between two reanalyses (Daget et al., 2009; Storto et al., 2013). In these studies, sea surface temperature (SST) has also been perturbed as a proxy for the heat fluxes using the same method. Milliff et al. (2011) and Pinardi et al. (2011) develop a Bayesian hierarchical model to generate surface vector wind and used them as perturbations for an ensemble system over the Mediterranean Sea.

In a different context, Wittenberg (2002), Harrison et al. (2002) and Zhang et al. (2005) develop a statistical atmospherical model to generate surface flux anomalies that they coupled with an ocean model to simulate and forecast ENSO events. The surface flux anomalies are generated by regression with respect to SST anomalies, and then possibly adding a stochastic part that is stationary in time. The regression weights are calculated using a singular value decomposition of the covariance matrix between SST and surface flux anomalies provided by different reanalyses.

In this article, we propose a model for generating global atmospheric forcing perturbations for each of the atmospheric variables used in the bulk formulae. Based on linear regression, this model generates spatially-varying multi-variate perturbations which are unbiased and whose spatial structure complies with the expected standard deviation of the true atmospheric variable errors. Once the model is designed, the perturbations are generated on- or off-line without relying on any external data and can be used globally without any regional limitations. These perturbations can then be applied on the atmospheric forcing variables themselves to seed an ensemble of any size more appropriately.

The paper is organised as follows. The design of the model is described in Section 2, from the training data to the method used to define the model parameters. In Section 3, the model is validated by looking at the distribution and the structure of the perturbations it generates. The study is then summarised and discussed in Section 4.

2. Model design

The model designed in this paper aims at providing atmospheric forcing perturbations for short and long wave radiation fluxes (swrd; lwrd), temperature and humidity at 2 m above the sea (t2m; q2m), zonal and meridional components of the wind at 10 m above the sea (u10m; v10m), precipitation and snow. These are the usual atmospheric parameters used in bulk formulae such as the CORE ones (Large and Yeager, 2004). The perturbations are meant to be of a random nature and should not include any biases. The model is based on linear regression in order to include some balance between the different variable perturbations.

2.1. Training data

The perturbation model has to be trained by a set of data that represent the error of the atmospheric forcing fields. For the CMCC C-GLORS reanalysis system (Storto et al., 2016) that is used in this study, the atmospheric forcing fields required by the bulk formulae according to Large and Yeager (2004), are provided by the ERA-Interim reanalysis as daily means for swrd and lwrd, precipitation and snow, and three-hourly means for t2m, q2m, u10m and v10m. A proxy of the atmospheric forcing error is obtained by calculating the difference between the ERA-Interim reanalysis and the MERRA reanalysis

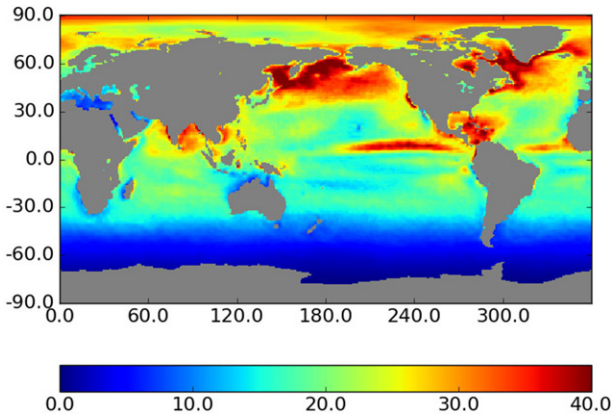

Fig. 1. Standard deviation of the reanalysis difference for swrd in June, July and August.

(Rienecker et al., 2011). Because the MERRA reanalysis has a higher resolution (0.5×0.67 degree), it is interpolated on the ERA-Interim regular grid 0.75×0.75 degree in order for the generated perturbations to be on the same grid as the fields they are perturbing. A ten-year period of this difference, from 2004 to 2013 is selected. In order to not account for possibly different diurnal cycles between the reanalyses, we are considering daily perturbations only, and hence, the three-hourly fields are daily averaged.

The variability of the reanalysis difference indicates for any locations how both reanalyses conform. It gives insights on the system differences due to parametrisations, assumptions, schemes for solving the equations, etc. As an example, Fig. 1 shows the standard deviation of the reanalyses difference for swrd in June, July and August. South of 50S, there is almost no variability, which is a consequence of the limited insolation received by the southern high latitudes during the austral winter. As expected, the highest variability (50 W/m^2) is found in the Northern hemisphere and the Equatorial regions. During the summer, the swrd is at its highest and day by day values can differ from one reanalysis to the other one because of disagreement in the cloud cover.

As illustrated in Fig. 1, the reanalysis difference variability varies depending on the regions. This variation differs from one variable to the other. It is hence difficult to define regions where a same model parametrisation could apply, and easier to consider all the locations (grid point) independently. For a particular variable at a particular location, the time series of the reanalysis difference shows as well some variations depending on the time of the year, requiring the model to account for these changes year-round. For example, Fig. 2a shows the training data for swrd at a high latitude location at 70N and 180E. The data have been sorted depending on the day of the year, i.e. the 1st January for the ten years, then the 2nd January for the ten years, etc. From mid-November to end of January, the reanalysis differences are all close to 0, but then increase to reach a range between –150 and 200 W/m^2. This example shows that a classical seasonal split is not the ideal solution, and a more proper technique for splitting the time is described in the next sections.

The time series of the reanalysis difference often includes a seasonal variation that must be removed to train the model at generating unbiased random perturbations. Classically, biases are removed by subtracting the mean calculated on a long enough period. Figure 3a shows the time series (solid black line) for lwrd at 87N and 4E. The sinusoidal shape of the time series is explained by the seasonal variation. If the yearly mean (dotted grey line) were to be used to unbias the data, it would shift the data without removing this seasonal variation. Therefore, the seasonal mean (dashed grey line) could be used instead. However, the anomalies with respect to the seasonal mean do not remove shorter time scale variations. In our example, this is particularly clear for autumn (Fig. 3b). A more efficient way for removing the seasonal variation is to high-pass filter the time series, in order to better select the different time scales to remove. We chose to apply a fourth-order Butterworth high-pass filter with a ten-day cut-off. The filter is applied twice to prevent any phase shift. As illustrated in Fig. 3c the filtered data do no longer have any seasonal variations. Since only the large scales are removed from the data, the variance of the raw data is almost entirely kept.

Once filtered, the reanalysis difference constitutes the training data that is used to design the model.

2.2. Independent variable

Designing a model using linear regression requires to define first an independent variable that can be modelled

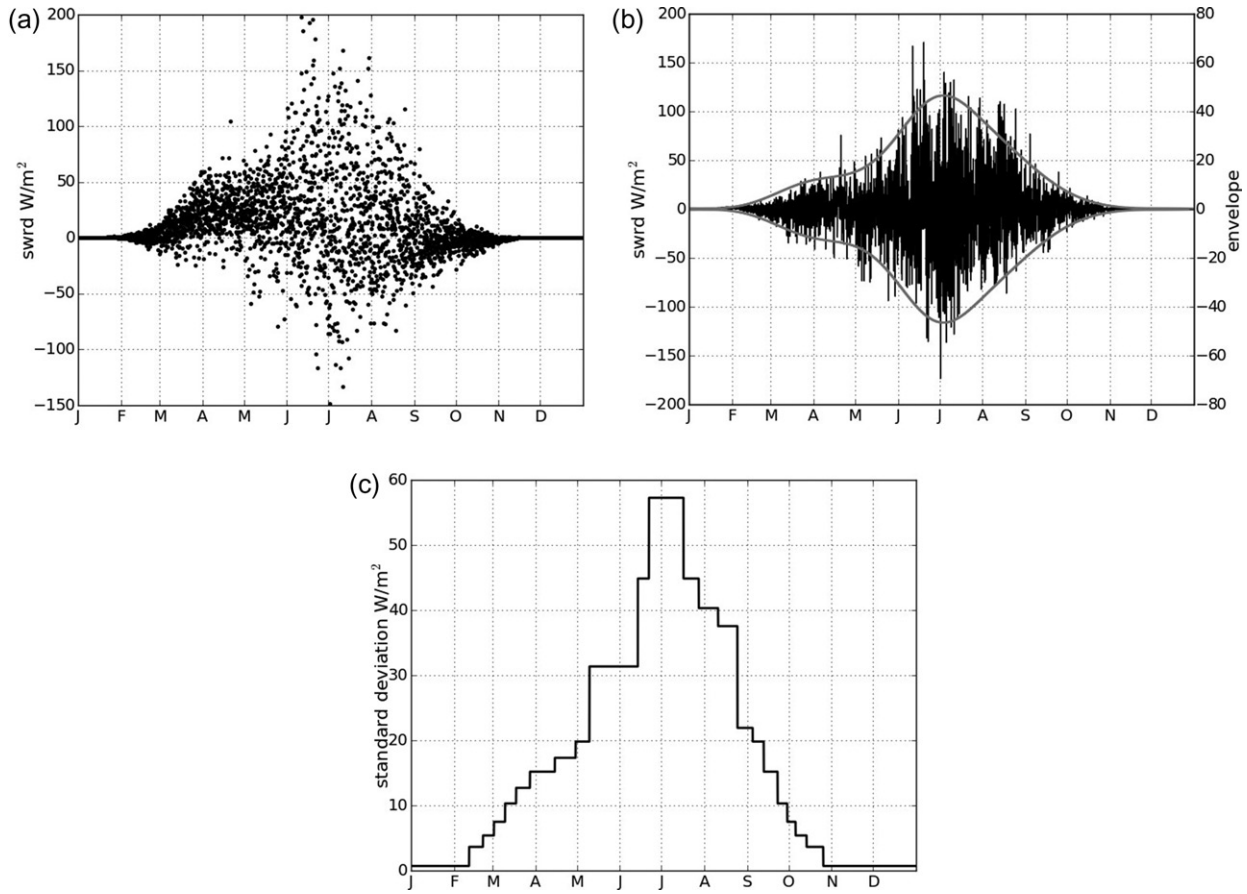

Fig. 2. Reanalysis differences for swrd at high latitude (70N, 180E). Panel (a) shows the raw data whilst panel (b) shows the filtered data (black line; left scale) along with the wave envelope (grey line; right scale). The ten-year data is sorted by days of the year. Panel (c) represents the standard deviation calculated for the model for each time of the year.

with a simple normal distribution. Swrd and lwrd have the training data whose distribution is closer to normal. Nevertheless, there exists a significant anti-correlation between them (not shown) and therefore only one should be selected as independent variable whilst the other one would be modelled from the former. Preliminary tests show that the validation (see Section 3) gives a slightly better performance when other variables are modelled with swrd rather than lwrd as an independent variable. Moreover, in the Earth energy budget, swrd can be viewed as an external source, although its location at the bottom of the atmosphere makes it depending on some atmospheric features such as clouds. Swrd is hence selected as the independent variable. At high latitudes and winter time however, because the values of swrd are close to 0, lwrd will mainly be modelled from its random component (see Section 2.3) as if it was an independent variable.

Modelling swrd through a normal distribution requires to define the standard deviation of the distribution. As shown in the example of Fig. 2b, the variation of the training data time series (black line) does not necessarily show a clear pattern that can be defined in a classical way (season, month, etc.). Following amplitude demodulation techniques, it is nevertheless possible to detect the wave envelope of the data, i.e. a signal representing the variation of the time series. To do so, the time series are firstly rectified (the negative values are replaced by their opposite) in order for the time series to have positive values only. A 4th order Butterworth low-pass filter is then applied with a five-day cut-off. The filter is applied twice to prevent any phase shift. Figure 2b shows the wave envelope (grey line; right scale) of the swrd time series at 70N and 180E. For a better perception, the opposite wave envelope has been plotted as well.

The wave envelope of a time series is a representation of the yearly evolution that the model standard deviation should follow but does not give the standard deviation directly. Nonetheless, it can be split into amplitude categories, and a standard deviation can be calculated for each one by gathering all the time series samples belonging to this category. The number of categories determines

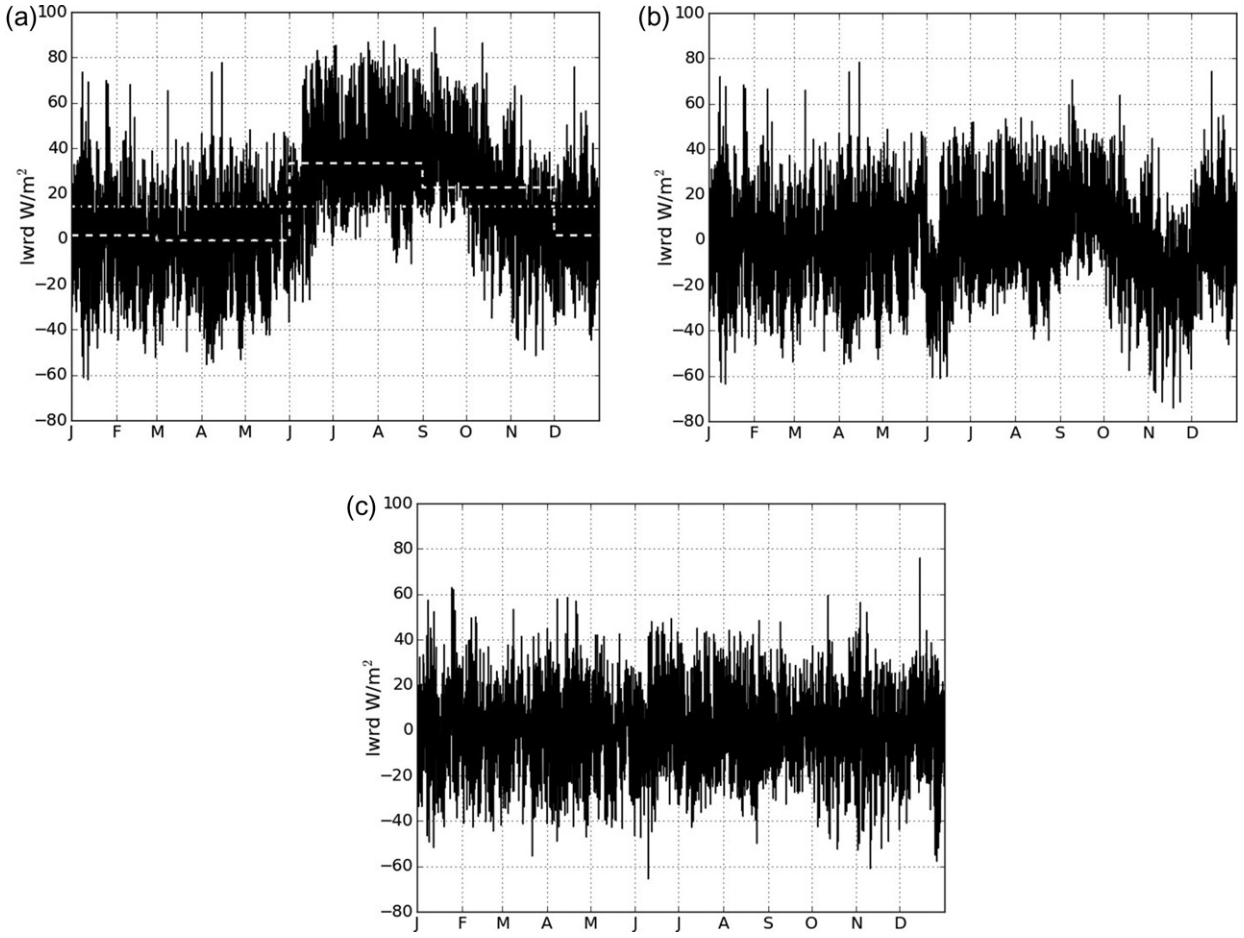

Fig. 3. Reanalysis differences (black lines) for lwrd at high latitude (87N, 4E). In panel a, the dotted grey line and the dashed grey line represent the annual and seasonal mean, respectively. The raw data (panel a) are filtered using the seasonal mean (panel b) or a high-pass filter (panel c).

the accuracy of the estimated standard deviation evolution. For example, splitting in two categories the wave envelope of Fig. 2b would lead to a standard deviation for the period May-August, and another one for the period September–April. For the latter, the modelled perturbations would be totally inaccurate in the period where they should be close to 0. On the other hand the calculation of the standard deviation must be done on a high enough number of samples to be robust. Moreover, since the modelling is done at each grid point, the more categories the model includes, the more model coefficients must be stored. In our case, we are looking for an accuracy that could validate our method, rather than the more efficient model. Therefore, we choose to split the wave envelope in a maximum of 25 categories. If a category contains less than five consecutive samples (strong increases or decreases of the wave envelope), the samples are transferred into the previous or following category.

For a particular category, the standard deviation is calculated by first gathering all the samples belonging to it, and checking the normality of their distribution. Following the philosophy of the D'Agostino-Pearson normality test (D'Agostino, 1970), a set of samples is considered normally distributed if its kurtosis and skew are within the range ±0.5. If the test fails, the presence of outliers is considered, and a maximum of 10% of the samples is allowed for rejection, before the standard deviation is actually calculated. Figure 2c shows the standard deviation for each time of the year that will be used for modelling swrd at the particular location 70N and 180E.

The model parameters to be stored for swrd consists in a pair of two-dimensional arrays (over the horizontal grid), one giving the time limits of the split for the year, and the other one giving the standard deviation to be applied for each of the splits.

2.3. Dependent variables

Apart from swrd, all the other variables are modelled through the sum of two components: a first component

Table 1. Sources for modelling each dependent variable. The same dependence relationships are applied for each grid point.

Variable	Source 1	Source 2
Long wave radiation flux (lwrd)	Short wave radiation flux (swrd)	
Zonal component of the wind (u10m)	Long wave radiation flux (lwrd)	
Meridional component of the wind (v10m)	Long wave radiation flux (lwrd)	
Temperature (t2m)	Long wave radiation flux (lwrd)	Short wave radiation flux (swrd)
Humidity (q2m)	Temperature (t2m)	Long wave radiation flux (lwrd)
Precipitation	Humidity (q2m)	Short wave radiation flux (swrd)
Snow	Precipitation	

depending on one or more other variables (called sources hereafter to avoid confusion), in order to introduce some balance between them; and a random component drawn from a normal distribution to represent the own variability of the variable:

$$\delta_{\text{variable}} = \sum_{i=1}^{N} \lambda_i \delta_{\text{source}_i} + \eta,$$

where δ is a perturbation, λ a regression coefficient and η a random component. For this study, the number of sources N is limited to two in order to keep the model simple enough. The sources of each dependent variable is chosen by calculating the correlation at each grid point between the dependent variable data and all the other variable data, and retaining then the ones showing patterns of significant correlations. For example, lwrd and swrd show a high temporal anti-correlation almost everywhere (about –0.6), whereas the correlation between lwrd and the other variables is smaller. Visualising spatially the temporal correlation is also useful to try and distinguish between the recursive correlations. For example, the spatial patterns of the temporal correlation between the components of the wind and swrd on the one hand, or lwrd on the other hand, are similar (not shown). But this similarity is probably due to the strong link that exists between swrd and lwrd, therefore only lwrd is retained as a source because of its slightly stronger correlation. For t2m however, the spatial patterns of the temporal correlation with lwrd on the one hand, and swrd on the other hand, differs, especially in the Tropics (not shown). Therefore, both lwrd and swrd are retained as sources for t2m. When there are two sources, they are sorted by their highest correlation. The sources for each dependent variables are chosen as shown on Table 1.

Once sources have been identified for a particular dependent variable, the associated regression coefficients (λ) are calculated depending on the correlation that exists between them. The correlation is calculated at each grid point for every day of the year. It is hence calculated from 10 values (10 years). A low-pass filter with a five-day cut-off is then applied twice (no phase shift) to smooth the correlation time series. As an example, Fig. 4 shows the correlation (solid line; left scale) between q2m as a dependent variable and t2m as a source in the Tropical Atlantic (6S, 30W). There is a strong anti-correlation during the austral spring that reaches –0.7 whereas during autumn, the relationship is much weaker. The correlation time series is then split into categories of amplitude 0.05. Once again, the number of categories determines the accuracy and the size of the parameters to store. If a category contains less than five consecutive samples (strong increases or decreases of the correlation time series), the samples are transferred into the previous or following category. For each categories, the data of the dependent variable and its source belonging to this category are then gathered to calculate a regression coefficient (λ). In our example Fig. 4, the dashed line (right scale) represents the regression coefficients depending on the time of the year for modelling q2m from t2m. When the correlation is weak, the regression coefficient is small, whereas when the correlation is strong, the regression coefficient is larger.

After the regression coefficients have been defined for a source, the part explained by this source is removed from the dependent variable data. If the variable has a second source the same process is followed to determine the regression coefficients for this second source, and the part it explains is also removed from the dependent variable data. The residual data are modelled by using a normal distribution and the standard deviations are defined as explained for the independent variable in the previous section. The model parameters to be stored for each dependent variable consists in a pair of two-dimensional arrays (over the horizontal grid) per source and a pair of two-dimensional arrays (over the horizontal grid) for the residuals. In each pair, the first array gives the time limits of the split for the year, and the other one gives either the regression coefficient or the standard deviation to be applied for each of the splits.

2.4. Generation of the perturbations

Once the model has been designed, the perturbations can be generated easily without the need of any other external

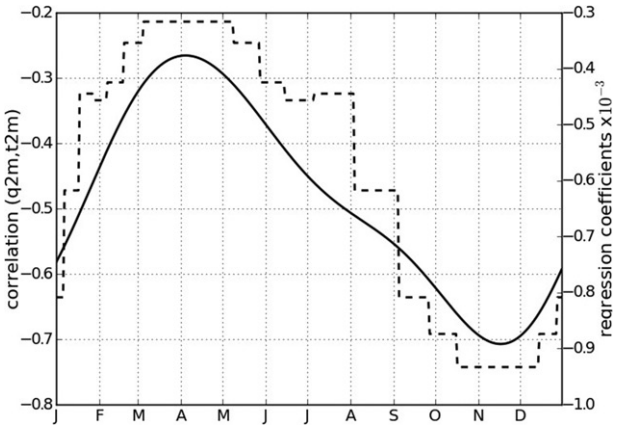

Fig. 4. Correlation between q2m and t2m (solid line; left scale) and the resulting regression coefficient (dashed line; right scale) for q2m at 6S and 30W.

data. Generating a set of perturbations for a specific day d of the year starts with generating the perturbations for the independent variable swrd

$$\boldsymbol{\delta}_{\text{swrd}} = \eta_{\text{swrd}} \boldsymbol{\Sigma}^{d}_{\text{swrd}},$$

where η_{swrd} is a random number drawn from a standard normal distribution (mean 0 and standard deviation 1), and $\boldsymbol{\Sigma}^{d}_{\text{swrd}}$ is the swrd standard deviation field for the specific day d. Using the same random number instead of drawing one for each grid point allows for the spatial structure to be preserved, in a global and not local sense. The other variables are generated according to their dependency: lwrd is generated from swrd, then the u10m and v10m are generated from lwrd, and so on. For example, the perturbations for t2m are generated as

$$\boldsymbol{\delta}_{\text{t2m}} = \boldsymbol{\Lambda}^{d}_{\text{t2m/lwrd}} \circ \boldsymbol{\delta}_{\text{lwrd}} + \boldsymbol{\Lambda}^{d}_{\text{t2m/swrd}} \circ \boldsymbol{\delta}_{\text{swrd}} + \eta_{\text{t2m}} \boldsymbol{\Sigma}^{d}_{\text{t2m}},$$

where $\boldsymbol{\Lambda}^{d}_{\text{t2m/lwrd}}$ and $\boldsymbol{\Lambda}^{d}_{\text{t2m/swrd}}$ are the regression coefficient fields of the specific day d for the sources lwrd and swrd, respectively, and $\circ$ represents a Schur product. As previously, η_{t2m} is a random number drawn from a standard normal distribution, and $\boldsymbol{\Sigma}^{d}_{\text{t2m}}$ is the t2m residual standard deviation field for the specific day d.

The generated perturbations can either be added or subtracted from the forcing fields. For semi-restricted value fields as swrd, lwrd, precipitation or q2m, which are lower bounded by 0, a truncation to ensure that their bounds are not violated is applied. When discussing their framework for generating perturbations for forcing fields, Turner et al. (2008) mention the detrimental effect that such a truncation can have on the distribution of the perturbations. To address the problem, their approach is to render the standard deviation linearly dependent upon the difference between the field value and the lower bound. The perturbation truncation that guarantees that no value will be negative is still required, although it does no longer affect the distribution significantly. For example, for a field like precipitation that has a large amount of zero-valued data, such a method prevents the introduction of large biases due to truncation. In our model, the problem is generally handled by the categorisation of the standard deviation. During dry periods where the values are small, the standard deviation is generally small, and the truncation does not affect the distribution significantly.

3. Model validation

The model is validated by comparing three years of generated perturbations to three years of original data. The original data are constituted of filtered differences between ERA-Interim and MERRA reanalyses as for the training data but for the years 2001 to 2003, which corresponds to the years before the training data period. Thus, they represent an independent dataset to validate the model against.

3.1. Distribution

The generated and original data are compared through a Kolmogorov–Smirnov test (Smirnov, 1939) that measures the distance between their empirical distribution functions. If the maximal distance is smaller than the critical distance, the null hypothesis H_0 = *'The samples are of the same distribution'* cannot be rejected. The critical distance depends on the number of samples and on a significance level α that is generally chosen as $\alpha = 5\%$. The Kolmogorov–Smirnov test is a non-parametric test, i.e. it does not make any assumptions on the distributions that are compared, and is sensitive to both location and scale. Figure 5 shows the results of the test for the different variables, with a significance level of $\alpha = 5\%$. In these plots,

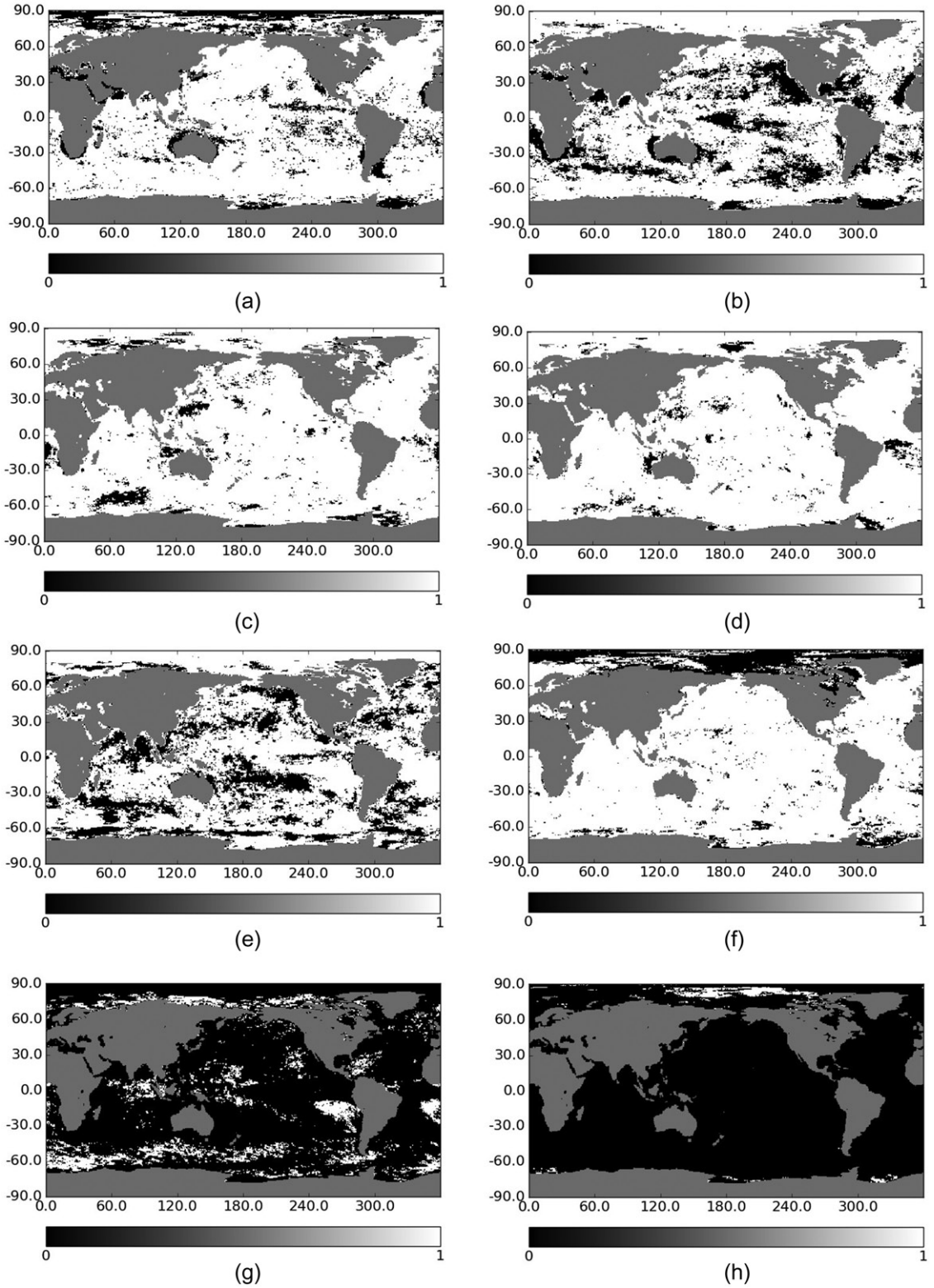

Fig. 5. Kolmogorov-Smirnov test result with $\alpha = 5\%$ for swrd (a), lwrd (b), u10m (c), v10m (d), t2m (e), q2m (f), precipitation (g) and snow (h). The null hypothesis is rejected for the grid points with values 0 (black), but cannot be rejected for the grid points with values 1 (white).

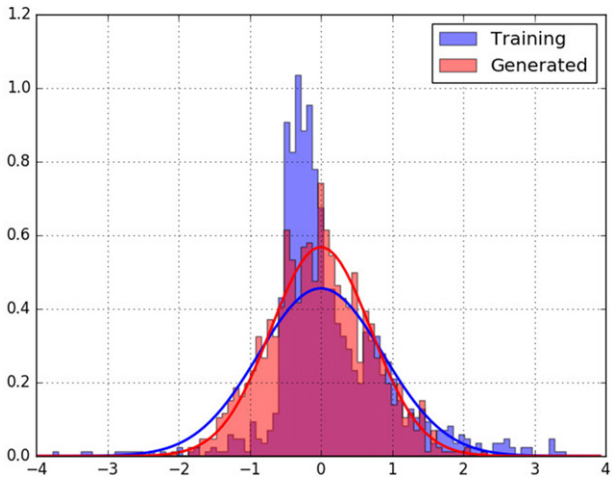

Fig. 6. Normed histogram for the training data (blue) and the generated data (red) for swrd at high latitude (70N, 180E) from end October to mid-February. The blue and red lines represent the theoretical normal distribution associated with the standard deviation of the training and generated data, respectively.

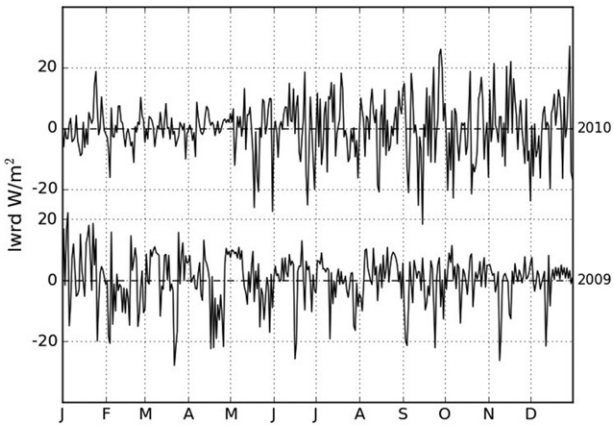

Fig. 7. Training data for lwrd for the years 2009 and 2010 in the equatorial Pacific (0N, 180E).

values at 0 (black) indicate regions where the null hypothesis has been rejected, whereas values at 1 (white) indicate regions where the null hypothesis cannot be rejected.

The independent variable swrd (Fig. 5a) shows a general good fit between the distributions except for the high latitudes in the Northern hemisphere. As mentioned before, the time series in these regions present null values in winter. This leads to a non-normal distribution with a high kurtosis. For example at the same location as in Fig. 2, the training data from end of October to mid-February has a kurtosis above 8, whereas the kurtosis for a normal distribution is about 0 (Fisher's definition). For this location and this period of the year, Fig. 6 shows the normed histogram of the training data (blue) where the high kurtosis and a right skew can be seen clearly. The normed histogram of a ten-year generated data is also shown (red). It is clear from these histograms that the distributions do not fit. This mismatch affects the distribution of the whole data set and leads the Kolmogorov–Smirnov test to fail. It is interesting to note the presence of outliers in the training data such that the standard deviation calculated from the whole data is smaller than the one calculated from the data where outliers have been rejected (see Section 2.2). This is illustrated on the figure from the normal distributions associated with the smallest standard deviation (blue line) or the highest standard deviation (red line). The latter corresponds to the distribution of the generated data.

For dependent variables, the problem of modelling a non-normal distribution from a normal distribution can also occur for the residuals. This is the case for example for q2m at high latitudes in the Northern hemisphere, with a very low q2m occurring during winter. For precipitation and snow the problem is emphasised and occurs in all regions. Indeed, the main differences from one reanalysis to another one come from front positions

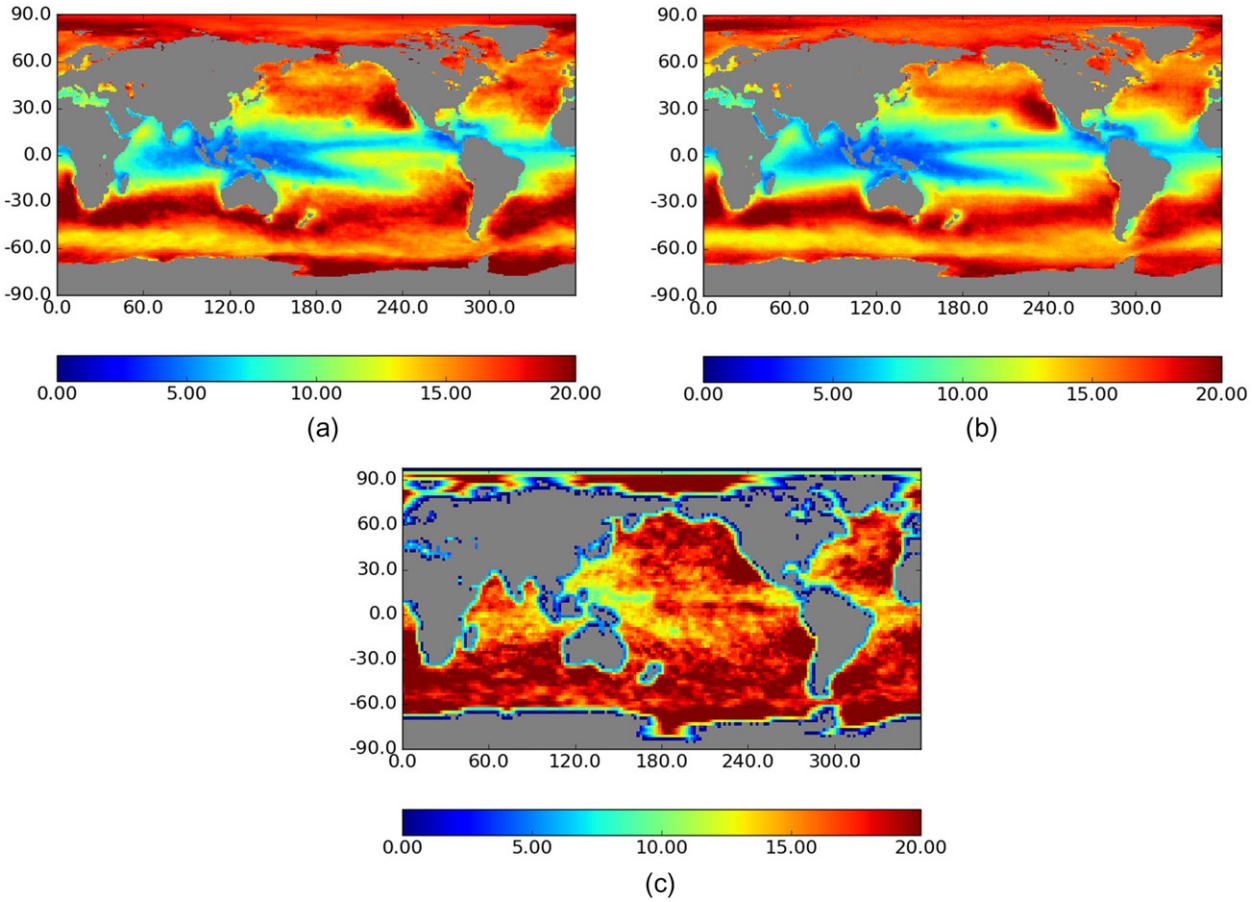

Fig. 8. Standard deviation of lwrd for three years of original data (a), perturbations generated by the model (b), perturbations generated from the previous CMCC method (c).

and precipitation intensity. The training data do generally not show a clear pattern during the year, but small differences with peaks here and there. The design of our model is clearly not adapted for precipitation and snow perturbations and therefore, the null hypothesis of a good fit between the distributions is rejected almost everywhere in the Kolmogorov–Smirnov test (Fig. 5g,h).

Whilst the Kolmogorov–Smirnov test shows a general good fit for u10m and v10m (Fig. 5c,d), the results are more mitigated for lwrd (Fig. 5b) and t2m (Fig. 5e). As mentioned before, the patterns in the training data from one year to the other can change drastically for some regions. For example, Fig. 7 shows the years 2009 and 2010 of the training data for lwrd in the Equatorial Pacific (0N, 180E). These two years are particularly different. From January to April, the reanalysis differences are in the range [–20, 20] W/m^2 in 2009 but only [–10, 10] W/m^2 for 2010. For the rest of the year the opposite occurs, with larger differences in 2010 than in 2009. The model cannot reproduce such differences between years in the perturbations. It produces only 'the best fit to all years' of the training data. If the original data present years so different, the Kolmogorov–Smirnov test will then fail.

Generating and comparing different three-year sets of perturbations does not lead to the exactly same results. For some regions the results can either be worsened or improved, and Fig. 5 shows intermediate results. The test is also sensitive to the number of years used in the samples. Using more or less years does not necessarily improve the results. This tends to confirm the issue of having different patterns in different years. An interesting fact is that q2m, that is modelled from t2m and lwrd, shows a good fit even in regions where results for its sources are not good. Failing the Kolmogorov–Smirnov test is hence not necessarily crippling, but gives insights on possible issues for particular regions and variables.

3.2. *Spatial structure*

The atmospheric forcing fields are the output states of an atmospheric system. Therefore, the error that affects them has a covariance structure. One of the aims of our

model is to try and represent some spatial structure in the generated perturbations. Although this structure, as the error, is unknown, a good indication is given by the patterns shown by the standard deviation of the training data. For each variable, the spatial patterns of the standard deviation are similar for the three years of original data and the three years of generated data. For example, Fig. 8a,b compares the standard deviation for lwrd. We can see clearly the similarity of the patterns, in particular the smallest variation at the Equator and the West Pacific warm pool. It is interesting to note that even for precipitation (not shown), a variable that fails the Kolmogorov–Smirnov test, the patterns are similar as well, although the variability is smaller than for the original data.

Figure 8c shows as well the standard deviation of three years of generated data but with the method used so far at CMCC (Storto et al., 2013). In this method, differences between the ERA-Interim and the NCEP/NCAR reanalyses are calculated and unbiased using the monthly mean. For each month, the daily perturbations are chosen randomly between the differences of this month. Some temporal correlation is then added using a first order autoregressive function. Note that these perturbations are on a coarser grid than the ERA-Interim one. With this method the variability of the generated perturbations is much too high, and it is difficult to see the patterns.

The model we designed succeeds at representing a consistent spatial structure in the generated perturbations, whereas the method used so far gives a too high variability.

4. Summary and discussions

A model to generate perturbations for atmospheric forcing fields has been designed, based on the ten-year (2004–2013) differences between the ERA-INTERIM and MERRA atmospheric reanalyses. Constructed from linear regressions, it aims at generating multi-variate unbiased perturbations with a consistent spatial structure for all the atmospheric variables required for the BULK formulae calculation (short and long wave radiation, components of the wind, temperature, humidity, precipitation and snow). The perturbations are directly generated on the atmospheric forcing grid used in the system, and hence, no interpolation is required. We chose in this study to use daily perturbations in order not to take into account any diurnal variation, although this could be envisaged in the future, owing to the increase of temporal resolution of atmospheric reanalysis products.

The short wave radiation (swrd) is selected as the independent variable, modelled solely from a normal distribution. From there, the other variables are modelled using one or two sources (other variables used as predictors) and a normal distribution for the residual. Dependencies are hence introduced between the different perturbations, mimicking the balance relationships existing between the atmospheric variables. The dependence relationships between the variables are fixed for all the grid points in this study. It could be interesting however, to allow for these relationships to be location-dependent. Moreover, the swrd perturbations are generated by drawing a single random number from a standard normal distribution, that is then multiplied by the proper standard deviation of the grid points, instead of drawing a random number for each grid point. Residuals for the dependent variables are modelled the same way. The spatial structure of the training data variability is hence preserved globally within the perturbations. This is confirmed by the standard deviation of the generated perturbations.

The standard deviation for swrd and for the other variable residuals was defined by categorising the envelope of the training data time series at each grid points. A similar method was applied to define the regression coefficients by categorising the envelope of the daily correlation existing between a variable and its source. The number of categories used in the definition of the parameters determines the accuracy and the cost of the model. The more categories, the better the representation, but also, the more parameters to store. In this study, the aim was to demonstrate the usefulness of such a model, and therefore a high number of categories was selected for all variables. To determine the optimal number of categories for each variable it would be recommended to run sensitivity tests.

Particular attention was paid to ensure that the training data was unbiased, and that no trend would exist that could be reproduced by the model. To perform this task, a high-pass filter was preferred to the classical seasonal or monthly mean removal. Moreover, the model parameters for each variable and at each grid point (standard deviations and regression coefficients) were defined by categorising the envelope of the time variations, allowing for more consistent time splits than the usual seasonal or monthly ones. For bounded variables such as radiation fluxes and humidity for example, confining the standard deviation in periods when the perturbation values are near the bound, limits any truncation and hence, restricts any associated changes of the distribution.

The model was validated by comparing the distributions of three-year generated perturbations against three years of unbiased differences between ERA-INTERIM and MERRA reanalyses, outside the training data period (2001–2003). A Kolmogorov test that compares the distributions of two samples was applied, and shows a reasonably good fit for all the variables except for precipitation and snow. The model is particularly in difficulty in regions where the training data presents a high kurtosis. This is especially true for precipitation and snow, but also

at high latitudes for swrd for example. For such features, modelling perturbations from a normal distribution is clearly inadequate, and the strategy should be rethought. A possible lead could be the use of anamorphosis to transform the non-Gaussian distributions (Brankart et al., 2012). In some cases, the model does not perform well because no clear pattern can be extracted from the training data. Getting a longer set of training data could possibly help at defining patterns to train the model.

The standard deviation of both three-year sets was also compared. They show similar patterns, indicating that the generated perturbations are well structured with respect to the expected real error structure. This is true for all variables, including those which failed the previous Kolmogorov–Smirnov test. The same comparison with three-year perturbations generated by the former method used at CMCC shows that the variability of the generated perturbations is so high that the patterns cannot be seen. In terms of structure, the model designed in this paper is a clear improvement.

With the method proposed in this paper, perturbations for the atmospheric forcing fields can be generated on demand in a flexible and cheap way. Once the model has been designed, the generation does not depend on any external data, and hence does not rely on the production of atmospheric reanalyses or forecasts. The generated perturbations are realistic and are spanned in the space defined by the training data, unlike other meaningful methods such as multivariate EOFs for instance, which perturbations are spanned in a more limited space given by the number of EOFs used. The perturbations generated in the model are daily means and are hence less dependent on temporal correlations than three-hourly means for example. However, a possible improvement of the method could include some temporal correlation between the perturbations generated for the independent variable swrd.

The atmospheric forcing perturbations generated by the model designed in this study have proven satisfying enough to be used in Observing System Simulation Experiments (OSSEs) using ensembles in the framework of the AtlantOS project. Results from these experiments will be soon reported in another article.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This project has received funding from the European Union as Horizon 2020 research and innovation programme under grant agreement No 633211

References

Alves, O. and Robert, C. 2005. Tropical Pacific Ocean model error covariances from Monte Carlo simulations. *Q. J. R. Meteorol. Soc.* **131**, 3643–3658. doi:10.1256/qj.05.113

Blockley, E. W., Martin, M. J., McLaren, A. J., Ryan, A. G., Waters, J. and co-authors. 2014. Recent developments of the Met Office operational ocean forecasting system: an overview and assessment of the new Global FOAM forecasts. *Geosci. Model Dev.* **7**, 2613–2638. doi:10.5194/gmd-7-2613-2014

Brankart, J. M., Candille, G., Garnier, F., Calone, C., Melet, A. and co-authors. 2015. A generic approach to explicit simulation of uncertainty in the NEMO ocean model. *Geosci. Model Dev.* **8**, 1285–1297. doi:10.5194/gmd-8-1285-2015

Brankart, J. M., Testut, C. E., Béal, D., Doron, M., Fontana, C. and co-authors. 2012. Towards an improved description of ocean uncertainties: effect of local anamorphic transformations on spatial correlations. *Ocean Sci.* **8**, 121–142. doi:10.5194/os-8-121-2012

Brusdal, L., Brankart, J. M., Halberstadt, G., Evensen, G., Brasseur, P. and co-authors. 2003. A demonstration of ensemble-based assimilation methods with a layered OGCM from the perspective of operational ocean forecasting systems. *J. Mar. Syst.* **40–41**, 253–289. doi:10.1016/S0924-7963(03)00021-6

Buehner, M. 2005. Ensemble-derived stationary and flow-dependent background-error covariances: evaluation in a quasi-operational NWP setting. *Q. J. R. Meteorol. Soc.* **131**, 1013–1043. doi:10.1256/qj.04.15

Clayton, A. M., Lorenc, A. C. and Barker, D. M. 2013. Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Q. J. R. Meteorol. Soc.* **139**, 1445–1461. doi:10.1002/qj.2054

Daget, N., Weaver, A. T. and Balmaseda, M. A. 2009. Ensemble estimation of background-error variances in a three-dimensional variational data assimilation system for the global ocean. *Q. J. R. Meteorol. Soc.* **135**, 1071–1094. doi:10.1002/qj.412

D'Agostino, R. B. 1970. Transformation to normality of the null distribution of g1. *Biometrika* **57**, 679–681.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P. and co-authors. 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137**, 553–597. doi:10.1002/qj.828

Evensen, G. 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* **99**, 10143–10162. doi:10.1029/94JC00572

Evensen, G. 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dyn.* **53**, 343–367. doi:10.1007/s10236-003-0036-9

Harrison, M. J., Rosati, A., Soden, B. J., Galanti, E. and Tziperman, E. 2002. An evaluation of air-sea flux products for ENSO simulation and prediction. *Mon. Weather Rev.* **130**, 723–732. doi:10.1175/1520-0493(2002)130<0723:AEOASF>2.0.CO;2

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D. and co-authors. 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

Large, W. G. and Yeager, S. G. 2004. Diurnal to decadal global forcing for ocean and sea-ice models: the data sets and flux climatologies, NCAR/TN-460, Boulder, USA.

Lorenc, A. 2003. The potential of the ensemble Kalman filter for NWP - a comparison with 4D-Var. *Q. J. R. Meteorol. Soc.* **129**, 3183–3203. doi:10.1256/qj.02.132

Milliff, R. F., Bonazzi, A., Wikle, C. K., Pinardi, N. and Berliner, L. M. 2011. Ocean ensemble forecasting. Part I: ensemble Mediterranean winds from a Bayesian hierarchical model. *Q. J. R. Meteorol. Soc.* **137**, 858–878. doi:10.1002/qj.767

Oddo, P., Storto, A., Dobricic, S., Russo, A., Lewis, C. and co-authors. 2016. A hybrid variational-ensemble data assimilation scheme with systematic error correction for limited-area ocean models. *Ocean Sci.* **12**, 1137–1153. doi:10.5194/os-12-1137-2016

Penny, S. G., Behringer, D. W., Carton, J. A. and Kalnay, E. 2015. A hybrid global ocean data assimilation system at NCEP. *Mon. Weather Rev.* **143**, 4660–4677. doi:10.1175/MWR-D-14-00376.1

Pinardi, N., Bonazzi, A., Dobricic, S., Milliff, R. F., Wikle, C. K. and co-authors. 2011. Ocean ensemble forecasting. Part II: mediterranean forecast system response. *Q. J. R. Meteorol. Soc.* **137**, 879–893. doi:10.1002/qj.816

Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J. and co-authors. 2011. MERRA: NASA's modern-era retrospective analysis for research and applications. *J. Clim.* **24**, 3624–3648. doi:10.1175/JCLI-D-11-00015.1

Smirnov, N. V. 1939. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscow.* **2**, 3–11.

Storto, A., Masina, S. and Dobricic, S. 2013. Ensemble spread-based assessment of observation impact: application to a global ocean analysis system. *Q. J. R. Meteorol. Soc.* **139**, 1842–1862. doi:10.1002/qj.2071

Storto, A., Masina, S. and Navarra, A. 2016. Evaluation of the CMCC eddy-permitting global ocean physical reanalysis system (C-GLORS, 1982-2012) and its assimilation components. *Q. J. R. Meteorol. Soc.* **142**, 738–758. doi:10.1002/qj.2673

Storto, A., Oddo, P., Cipollone, A., Mirouze, I. and Lemieux, D. B. 2018. Extending an oceanographic variational scheme to allow for affordable hybrid and four-dimensional data assimilation. *Ocean Model* **128**, 67–86. doi:10.1016/j.ocemod.2018.06.005

Storto, A. and Randriamampianina, R. 2010. Ensemble variational assimilation for the representation of background error covariances in a high-latitude regional model. *J. Geophys. Res* **115**, 1–21. doi:10.1029/2009JD013111.

Toth, Z. and Kalnay, E. 1997. Ensemble forecasting at NCEP and the breeding method. *Mon. Weather Rev.* **125**, 3297–3319. doi:10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2

Turner, M. R. J., Walker, J. P. and Oke, P. R. 2008. Ensemble member generation for sequential data assimilation. *Remote Sens. Environ.* **112**, 1421–1433. doi:10.1016/j.rse.2007.02.042

Whitaker, J. S. and Hamill, T. M. 2002. Ensemble data assimilation without perturbed observations. *Mon. Weather Rev.* **130**, 1913–1924. doi:10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2

Wittenberg, A. T. 2002. *ENSO response to altered climates*. PhD thesis, Princeton University.

Zhang, S., Harrison, M. J., Wittenberg, A. T., Rosati, A., Anderson, J. L. and co-authors. 2005. Initialization of an ENSO forecast system using a parallelized ensemble filter. *Mon. Weather Rev.* **133**, 3176–3201. doi:10.1175/MWR3024.1

Zupanski, M., Fletcher, S. J., Navon, I. M., Uzunoglu, B., Heikes, R. P. and co-authors. 2006. Initiation of ensemble data assimilation. *Tellus A* **58**, 159–170. doi:10.1111/j.1600-0870.2006.00173.x