Clim Dyn
DOI 10.1007/s00382-016-3407-x

# Climate model forecast biases assessed with a perturbed physics ensemble

David P. Mulholland[1] · Keith Haines[2] · Sarah N. Sparrow[3] · David Wallom[3]

Received: 27 April 2016 / Accepted: 15 October 2016

**Abstract** Perturbed physics ensembles have often been used to analyse long-timescale climate model behaviour, but have been used less often to study model processes on shorter timescales. We combine a transient perturbed physics ensemble with a set of initialised forecasts to deduce regional process errors present in the standard HadCM3 model, which cause the model to drift in the early stages of the forecast. First, it is shown that the transient drifts in the perturbed physics ensembles can be used to recover quantitatively the parameters that were perturbed. The parameters which exert most influence on the drifts vary regionally, but upper ocean mixing and atmospheric convective processes are particularly important on the 1-month timescale. Drifts in the initialised forecasts are then used to recover the 'equivalent parameter perturbations', which allow identification of the physical processes that may be at fault in the HadCM3 representation of the real world. Most parameters show positive and negative adjustments in different regions, indicating that standard HadCM3 values represent a global compromise. The method is verified by correcting an unusually widespread positive bias in the strength of wind-driven ocean mixing, with forecast drifts reduced in a large number of areas as a result. This method could therefore be used to improve the skill of initialised climate model forecasts by reducing model biases through regional adjustments to physical processes, either by tuning or targeted parametrisation refinement. Further, such regionally tuned models might also significantly outperform standard climate models, with global parameter configurations, in longer-term climate studies.

**Keywords** Perturbed physics · Model drift · HadCM3 · Climate model optimisation

## 1 Introduction

Model drift, the development of forecast error through the manifestation of systematic model biases, is an important component of numerical forecasts on a range of timescales (Vitart 2004; Jung 2005; Magnusson et al. 2013), and one which limits their usefulness (e.g., Ding et al. 2015). To an extent, the physical parameters used in weather and climate models are 'tuned' in order to try to minimise these drifts, and to make the models as realistic as possible, but the 'true' or optimum values for these parameters are in general unknown, due to the lack of a measurable physical equivalent, or inability to test numerically the full range of possible values of all parameters (Severijns and Hazeleger 2005; Randall et al. 2007). Additionally, for forecasts using models with high spatial resolution, performing a comprehensive multi-parameter tuning may be prohibitively expensive (Annan et al. 2005), while the tuning of parameters on an individual basis is problematic in practice due to the interactions that occur between parameters (Collins et al. 2011; Williamson et al. 2015).

Using lower-resolution coupled models, tuning procedures involving perturbed physics ensembles have previously been performed, targeting a realistic long-term model climate (e.g. Murphy et al. 2004; Knight et al.

✉ David P. Mulholland
d.p.mulholland@reading.ac.uk

Keith Haines
k.haines@reading.ac.uk

[1] Department of Meteorology, University of Reading, Reading, UK

[2] Department of Meteorology and National Centre for Earth Observation, University of Reading, Reading, UK

[3] Oxford e-Research Centre, University of Oxford, Oxford, UK

Published online: 26 October 2016

2007; Brierley et al. 2010; Collins et al. 2011). However, these same coupled models are now being used for initialised forecasts (e.g., MacLachlan et al. 2014), and the set of parameters resulting from climate tuning can lead to large regional drifts in such forecasts (where the model has assimilated full-field observational data) of length one month, or up to several years. At the same time, however, it has been argued that there should be some consistency between biases at short and long lead times (Rodwell and Palmer 2007; Ma et al. 2014), such that short-term drift could provide better metrics for tuning climate model processes. Model drifts from an initialised state should contain useful information about climate model biases, although it is important to bear in mind that drifts are often a non-linear function of forecast lead time (Alessandri et al. 2010; Doblas-Reyes et al. 2013) and may also be state- and/or seasonally dependent (Kumar et al. 2012; Vannière et al. 2013).

Our aim in this work is to show that perturbed physics methods can also be used to study transient model drifts. By imposing an instantaneous change in parameter values during a model simulation, the model has in effect been initialised with a state that does not lie on its attractor. The model will, therefore, subsequently drift towards the new attractor, and away from a control run defined as the continuation of the simulation without changes to the parameters. This situation is conceptually equivalent to the problem of initialised (real-world) forecasts, in which a model is initialised using observational information which, due to the presence of physical or dynamical process biases, puts the model in an off-attractor state. The initialised forecast will then drift back towards the model's own attractor, in this case deviating from the true (observed) state evolution. Drifts (i.e., either model minus control for perturbed physics, or model minus analysed truth for initialised forecasts) in these two situations can therefore be compared. Methods developed to recover parameter perturbations from drifts in the former case may then be used to make deductions about the physical causes of biases in the latter case.

The climate model and method used are described in Sect. 2. Results using a perturbed physics ensemble to test the ability of a statistical model to use drifts to recover unseen parameter deviations are presented in Sect. 3. The same statistical model is then applied to an initialised hindcast (historical re-forecast) set in Sect. 4. The potential use of the method and its limitations are discussed in Sect. 5, and the main results are summarised in Sect. 6.

## 2 Methods

The coupled atmosphere-ocean model HadCM3 (Gordon et al. 2000), with an atmosphere horizontal resolution of 3.75° × 2.5° with 19 vertical levels extending to 40 km, and an ocean horizontal resolution of 1.25° × 1.25° with 20 vertical levels, was used to perform two sets of model experiments.

### 2.1 CPDN ensembles

A perturbed physics ensemble of 500 HadCM3 members was run on the climate*prediction*.net (CPDN) platform, a facility which allows large numbers of climate model simulations to be performed by utilising latent computing time provided by members of the public (see e.g., Allen 2003; Piani et al. 2005; Sanderson et al. 2008; Yamazaki et al. 2013). Multiple physical parameters in both atmosphere and ocean were perturbed simultaneously in each member using combinations of parameters previously found to give plausible realisations of the climate, as measured by top-of-atmosphere (TOA) radiative flux balance (Yamazaki et al. 2013). The control member used the standard HadCM3 values (Gordon et al. 2000) for all parameters.

All members were run for one year, starting on 1 December, from the same initial conditions, obtained by integrating the control version of the model from the pre-industrial era through the twentieth century, using both natural and anthropogenic forcing from 1880 onwards. Note that the control run is not otherwise nudged to observations, so named years do not correspond closely to the real-world equivalents. The continuation of the control run acts as a baseline for assessing the perturbed physics drifts, since all other members experience an instantaneous change in parameter values at the beginning of the simulation, causing them to drift from the control member. This process was repeated for four different start years, dated as 1940, 1950, 1960, and 1980 (although not initialised from real-world conditions on these dates), forming an ensemble of 500 members per year, with the same parameter perturbations across the four years. These initial conditions were chosen to span the model's climate variability, in particular with respect to ENSO (see "Recovery sensitivity tests" in Appendix), as well as a range of external forcings. This ensemble is named PP_FULL.

In addition, to explore sensitivity to ensemble size, a larger ensemble of 2200 members was run from the '1980' initial conditions (named PP_1980), with perturbations applied to the same set of parameters but with a denser sampling of parameter space. Smaller test ensembles of 50 members each were also run (collectively named PP_IC) using dates from '1930' to '1990', separated by 10 years, to test initial condition sensitivity. These sensitivities are explained further in the next section. For clarity, the various groups of perturbed physics members are listed in Table 1.

**Table 1** List of runs performed using HadCM3

| Run name | Start dates | Members per year | Purpose |
|---|---|---|---|
| PP_FULL | *Dec 1 1940, 1950, 1960, 1980* | 500 | Final predictive model |
| PP_1980 | *Dec 1 1980* | 2200 | Preliminary model evaluation and Sensitivity tests for training set size |
| PP_IC | *Dec 1 1930, 1940, ..., 1990* | 50 | Initial condition sensitivity tests |
| INIT | Nov 1 1965, 1970, ..., 2000 | 1 | Initialised hindcasts |
| INIT_MOD | Nov 1 1965, 1970, ..., 2000 | 1 | Initialised hindcasts with Modified parameter MLLAM |

Each member ran for one year. Start years for the three non-initialised ensembles are italicised to denote the fact that these do not represent real-world conditions for those years, unlike those used by INIT and INIT_MOD

In total, 31 atmosphere and ocean model parameters were varied independently, using the space-filling Latin hypercube algorithm (Yamazaki et al. 2013). These parameters are listed in Table 2, along with a short description of the physical role of each. In many cases, values were sampled continuously from a range of plausible values, while in some cases (i.e., integer parameters) a set of discrete values were sampled. The ranges of most parameters extend both above and below the control value, but in a few cases only values either greater or smaller than the control were used, as shown in the Table 2. In all cases, normalised parameter perturbations, $\Delta$p, are defined by subtracting the control value from the perturbed value, and dividing by the control value (since all control values are positive). The range of $\Delta$p for each parameter is given in Table 2.

### 2.2 Initialised hindcasts

A set of initialised hindcasts, with 8 start dates at five-year intervals from 1965 to 2000, was performed (INIT; see Table 1), using the same model and parameters as the control member of PP_FULL. These hindcasts each ran for one year, with initial conditions taken from the DePreSys 'full field'-initialised decadal hindcasts presented in Smith et al. (2013), which are only available on 1 November each year. These analyses were derived by nudging HadCM3 towards full-field atmospheric (Uppala et al. 2005) and oceanic (Smith and Murphy 2007) reanalyses. This provided 8 distinct realisations of the model drift, as a result of the range of initial conditions sampling both natural climate variability and the increasing greenhouse gas loading trend. These 8 members were then combined to produce ensemble-average drifts.

Following the results of Sect. 4.1, a further set of initialised hindcasts were performed (INIT_MOD; see Table 1). These were identical to INIT except that the wind mixing energy scaling parameter, MLLAM, following its identification in the INIT drifts, was reduced from its standard value of 0.7 to 0.5. This experiment is explained in more detail in Sect. 4.2.

### 2.3 Calculation of drifts

For all runs, monthly mean model output was stored for a selection of atmosphere and ocean fields, obtaining measures of the drift in the system as a whole, but with some focus on the near-surface due to its importance in forecasting. The atmospheric fields used were: 1.5 m temperature, 1.5 m relative humidity, precipitation rate, 10 m zonal and meridional winds, mean sea level pressure, surface latent heat flux, 250 hPa zonal and meridional winds, and TOA net shortwave and longwave radiative fluxes. In the ocean, sea surface temperature, sea surface salinity, mixed layer depth and sea ice fraction fields were stored at all gridpoints. Additionally, in the tropics (20°N–20°S), 3-D temperature and salinity were output on 10 vertical levels extending down to 300 m depth (below which drifts in the first month may be relatively small). Subsurface ocean information outside the tropical latitudes was available only through the mixed layer depth field.

For the analysis presented below, all model output was converted into spatial averages in 65 geographic regions. The 22 regions from Giorgi and Francisco (2000) (defined to represent specific climatic regimes, primarily covering land but also extending over the ocean in most cases) were used, following Sanderson et al. (2008), and these were supplemented with 43 ocean regions to give near-global coverage (the regional boundaries can be seen in Fig. 3a). The intention is that regions should capture coherent drifts that can be mapped onto consistent physical process errors. Subsurface ocean fields were used only in regions for which at least half of the area falls within 20°N–20°S (a total of 22 regions).

All model output data were converted into drifts, or errors, by subtracting either the control member, in the case of the perturbed physics ensemble, or the analyses obtained from DePreSys (Smith et al. 2007) in the case of the initialised hindcasts. The perturbed physics ensemble and the initialised hindcast experiments are now conceptually equivalent, as both involve a model initialised in a state that does

**Table 2** List of the parameters varied independently across the perturbed physics ensemble, separated into those affecting atmospheric physics and dynamics (including sea ice), atmospheric chemistry, and ocean dynamics

| Name | Min $\Delta$p | Max $\Delta$p | Description |
|---|---|---|---|
| Atmospheric physics and dynamics | | | |
| VF1 | −0.69 | 1.35 | Ice particle fall speed |
| CT | −0.50 | 4.62 | Cloud liquid water-precipitation conversion rate |
| CW_LAND | −0.60 | 8.93 | Threshold cloud liquid water content, sea |
| CW_SEA | −0.50 | 8.93 | Threshold cloud liquid water content, land |
| RCRIT | −0.14 | 0.29 | Critical rel. humidity for cloud formation (levels 4–19) |
| EACF | 0.00 | 0.60 | Empirically adjusted cloud fraction |
| ENTCOEF | −0.78 | 1.98 | Atmospheric entrainment rate coefficient |
| ALPHAM | 0.00 | 0.30 | Albedo at sea ice melting point |
| DTICE | −0.80 | 0.00 | Temperature range over which ice albedo varies |
| ICE_SIZE | −0.33 | 1.67 | Ice particle size |
| START_LEVEL_GWDRAG | 0.00 | 0.67 | Lowest model level for gravity wave drag |
| KAY_GWAVE | −0.50 | 0.00 | Surface gravity wave drag wavelength |
| KAY_LEE_GWAVE | −0.50 | 0.00 | Surface gravity wave trapped lee wave constant |
| ASYM_LAMBDA | −0.92 | 3.06 | Vert. distance before air parcels mix with surround |
| CHARNOCK | 0.00 | 1.00 | Const. in Charnock formula (mom. transport over sea) |
| Z0FSEA | −0.84 | 3.77 | Sea surface roughness (heat, moisture transport) |
| G0 | −0.75 | 1.25 | Used in calc. of boundary layer stability function |
| R_LAYERS | −0.50 | 0.00 | Num. of soil levels from which water can be extracted |
| DIFF_COEF | −0.88 | 0.25 | Horizontal diffusion coefficient |
| DIFF_EXP | −0.33 | 0.00 | Order of horizontal diffusion |
| Atmospheric chemistry | | | |
| L0 | −0.67 | 2.00 | Sulphate mass scavenging parameter L0 |
| L1 | −0.67 | 2.00 | Sulphate mass scavenging parameter L1 |
| SO2_HIGH_LEVEL | 0.00 | 0.67 | Model level for $SO_2$ (high level) emissions |
| VOLSCA | −0.50 | 2.49 | Scaling factor for volcanic emissions |
| ANTHSCA | −0.75 | 0.75 | Scaling factor for anthropogenic sulphate aerosols |
| Ocean dynamics | | | |
| ISOPYC | −0.90 | 1.50 | Surface tracer isopycnal diffusion |
| VERTVISC | −0.89 | 9.98 | Background momentum vertical diffusion |
| VDIFFSURF | −0.89 | 2.09 | Background vertical diffusion of tracer at surface |
| VDIFFDEPTH | −0.74 | 2.50 | Increase of tracer background diffusion with depth |
| MLLAM | −1.00 | 1.14 | Wind mixing energy scaling factor |
| MLDEL | −1.00 | 0.50 | Wind mixing energy rate of decay with depth |

Deviations from the control value are divided by the control value to produce normalised deviations $\Delta$p, given here to two decimal places

not lie on the model attractor, which subsequently drifts back towards that attractor. In the case of the hindcasts, the drift occurs due to the presence of physical or dynamical biases that prevent the model from accurately simulating the analysed ‘truth’, while in the case of the perturbed physics ensemble, the altered parameters cause the model to drift from the control member that represents the truth. The equivalence of these two sets forms the basis of this work, and allows us to make deductions about drifts/errors in initialised hindcasts through comparison with the CPDN ensemble.

## 3 Perturbed physics results

### 3.1 Model drifts

Illustrative examples of the drifts in sea surface temperature (SST) and precipitation in the western tropical Indian ocean, for both the perturbed physics ensemble (using PP_FULL) and the hindcasts (INIT), are shown in Fig. 1. In each case, drifts are normalised by dividing by the standard deviation of the 8 monthly mean DePreSys analysis values at the appropriate verifying month, which exists due to both

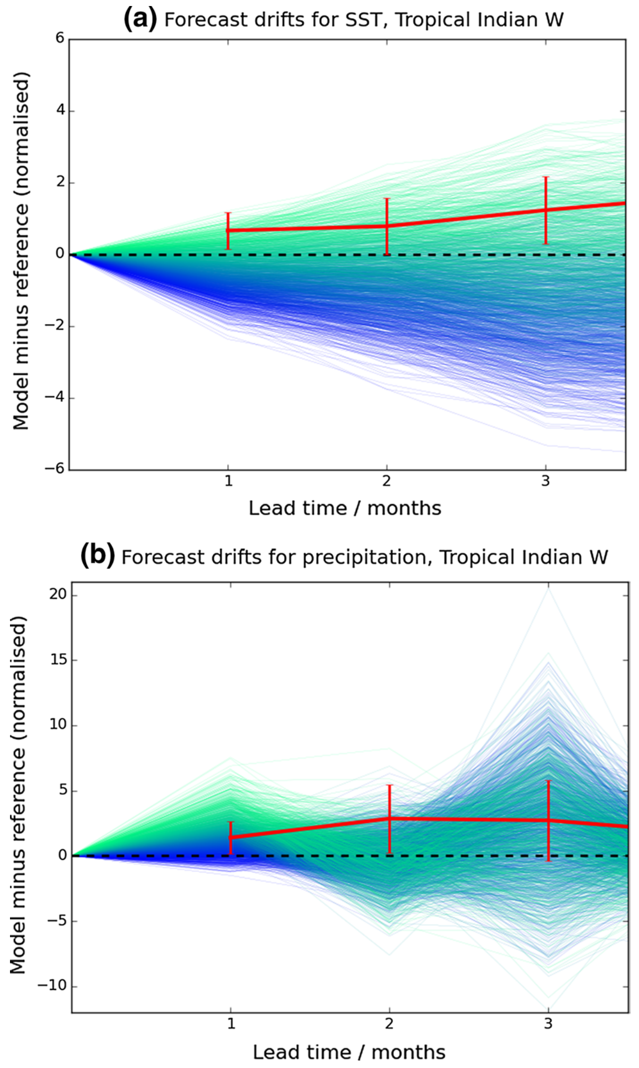

**Fig. 1** Examples of perturbed physics drifts from the 2000 members of PP_FULL, normalised by dividing by the standard deviation of the 8 DePreSys analysis monthly mean absolute values at the appropriate lead time, for **a** SST and **b** precipitation rate, in the western part of the tropical Indian basin. *Lines* are *coloured* by their value at month 1. The mean drifts from the initialised hindcasts (INIT), normalised in the same way, are *plotted in red* for comparison, with the standard deviation of the 8 normalised hindcast drifts shown by the *error bars*

'weather noise' and climate variability. The parameter perturbations are usually large enough that they produce drifts that are comparable to this spread (normalised values of order one).

For SST, ensemble drifts are generally consistent over the first few months, indicating that month-1 drifts capture well the dominant physical processes. However, for precipitation, there is strong month-to-month variability, as initial drifts occur more rapidly (within the first month). Drifts in ocean variables may often be more easily linked to a specific physical cause, driven by changes in one or several parameter values, while a one-month lead time may be too long to robustly link some atmospheric drifts to parameter changes, particularly in midlatitudes. Nonetheless, it was found that drifts in all fields (both ocean and atmosphere) do contribute to the recovery of parameter perturbations from model drifts, as described in the next subsection.

### 3.2 Estimating known parameter perturbations from model drifts

We test a method of using drifts to estimate the parameter perturbations in unseen perturbed physics ensemble members. The results of these tests also define methods to assign confidence levels to the inferences that are made about the initialised hindcasts in Sect. 4.

Drifts are obtained from all available model fields for the first month of a training set of simulations, along with their known parameter perturbations (in their normalised forms), and these are used as inputs to train a statistical inference model. This is then used to estimate the parameter perturbations from other 'unseen' perturbed physics simulations, based on their first-month drifts. A neural network was used for this task, although any of several common supervised machine learning algorithms could alternatively be used. Some details on this procedure are provided in Appendix "Statistical model of parameter-drift relationships". The procedure was performed separately for each region and for each parameter: although the perturbed physics ensemble uses global parameter changes only, the hindcast drifts are caused by regionally dependent process errors, such that regional recovery is necessary.

The basis of this method is the idea that a perturbation to a particular parameter produces a 'drift signature': correlated drifts in multiple model variables in the region of interest. For example, in the East Africa region, which includes part of the western Indian Ocean, a 10% increase in the atmospheric entrainment rate reduces the depth of atmospheric convection, causing large drifts (up to 20% of weather/climate variability) in several related fields, including decreased precipitation, increased surface pressure, and increased TOA shortwave radiation due to decreased cloud cover. Subsurface ocean temperature and salinity profiles also drift in a coherent manner, with opposite signs in the mixed layer and at depth in the case of temperature. The trained model extracts this signature from the test cases to determine the extent to which each parameter may have been perturbed to produce the observed drifts.

The performance of the neural network model was tested quantitatively using PP_FULL, split into training and test sets. Three start dates were chosen at random, and these 1500 drift signatures were used as the training set, with the remaining date used as the test set. Predicted

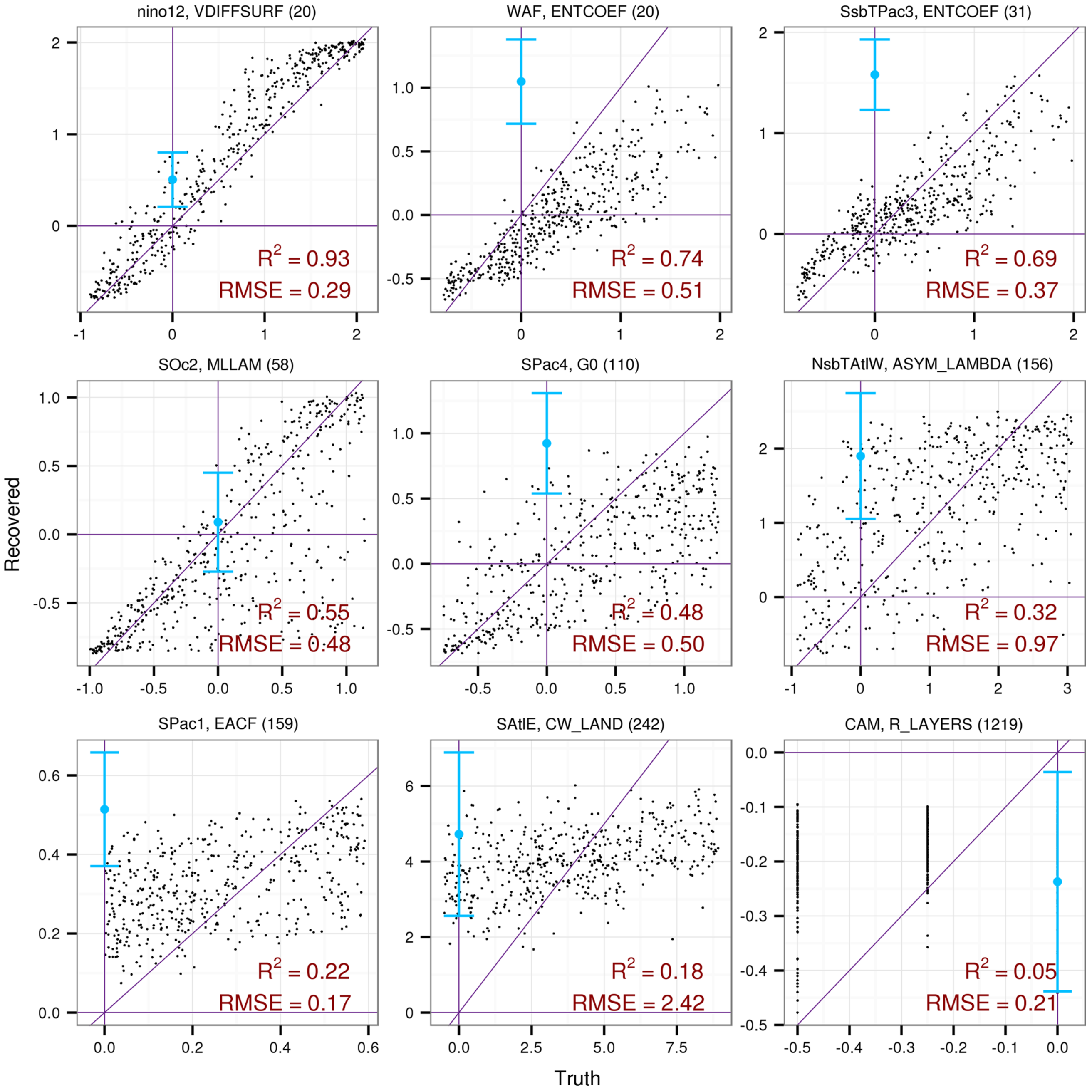

**Fig. 2** Examples of parameter recovery for each of nine $R^2$ quantiles, decreasing from 0.8–1 (*top left*) by decile to 0–0.1 (*bottom right*): predicted $\Delta$p values are scattered against the true values, for the 500 '1960' members of PP_FULL, which are used as the test set for these examples (the training set consisted of the other 3 years: '1940', '1950', and '1980'). The $R^2$ and RMSE values for the group of recoveries are printed in each plot, and the *number in brackets* in the title of each *panel* is the overall number of parameter-region combinations falling in that recovery accuracy $R^2$ quantile. The *blue points* show the value predicted for INIT (see Sect. 4), with the true RMSE (calculated as the average of the four possible training/test splits) shown as the *error bar* on this estimate. See Fig. 3a for region definitions

test perturbation values $\Delta$p were then compared to the true values for all 500 test members, and performance measured using squared correlations ($R^2$) and root-mean-square errors (RMSEs). This process was repeated with each start date acting as the test set, and the resulting metrics were averaged.

Examples of parameter perturbation $\Delta$p recovery using a single test year, the '1960' members of PP_FULL, are shown, for several parameters and regions, in Fig. 2 (recovered $\Delta$p values on the vertical axes, true values on the horizontal axes). In the best cases, parameter values can be recovered very accurately [e.g., vertical tracer diffusion

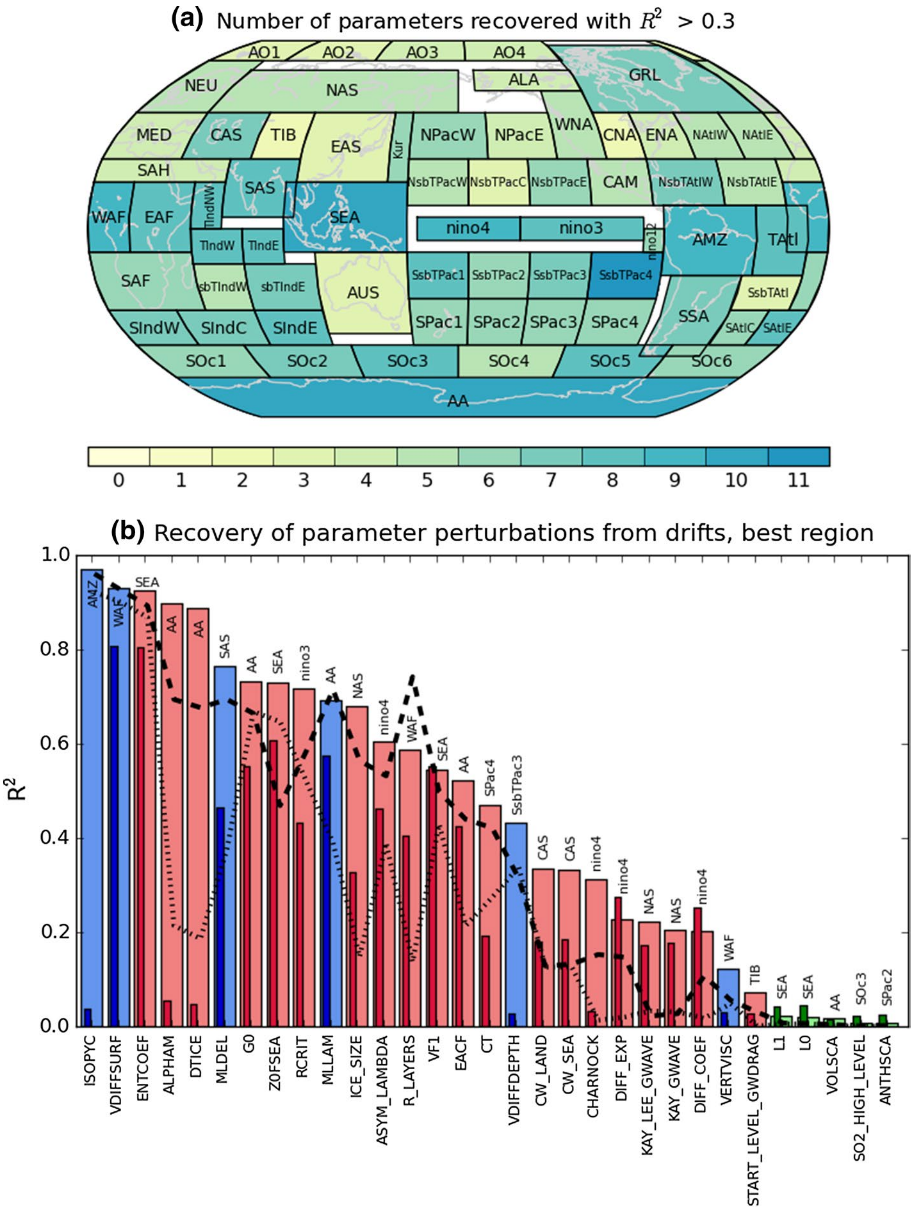

**Fig. 3** Validation of the parameter recovery method using PP_FULL, as an estimate of the skill of the final form of the predictive model. Month-1 drifts from three of the four start dates (1500 members) are used to estimate the parameter perturbations from the remaining start date. Shown are **a** the number of parameters that can be estimated with an accuracy of $R^2 > 0.3$ in each region; **b** all parameters ordered by how accurately they can be recovered, using as a measure of this the largest $R^2$ found for each parameter (*wide*, *pale bars*), along with the region in which this is obtained. In **b**, *bars* are *coloured* by parameter type (physical atmosphere (including sea ice) in *red*, ocean in *blue*, chemistry in *green*); the *dashed (dotted) line* shows the equivalent calculation for month-2 (6) drifts; and the *thin, dark bars* show the equivalent month-1 calculation using globally averaged drifts

(VDIFFSURF) in the Niño1–Niño2 region (nino12)] resulting in a high $R^2$ and a low RMSE (in normalised units) for the verification of the estimates. Other parameters, such as the boundary layer stability parameter G0 in the southern Pacific (SPac4), have a weaker impact on drifts, and can only be recovered with modest skill in most regions. It can be seen from these examples that an $R^2$ of ~0.3–0.5, while still associated with a large spread of $\Delta$p estimates, may still permit the perturbation sign to be identified (e.g., note the fraction of points lying in the upper-right and lower-left quadrants for G0 in SPac4).

Some parameters, such as ice albedo temperature range (DTICE), affect drifts only in very localised areas, and perturbations can only be recovered in one or a few regions, but an accurate recovery may nonetheless be possible from these relevant regions (see further recovery examples in Appendix "Statistical model of parameter-drift relationships"). Finally, some parameters cannot be recovered with useful skill in any region (e.g., chemistry parameters such as sulphate aerosol scaling factor (ANTHSCA), not shown), as would be expected. Since the set of parameters being perturbed was originally chosen in regard to climate sensitivity, it is not surprising that a subset has little effect on short-range simulations.

The RMSE values calculated from these recovery tests correspond to the vertical distances of points from the 1:1 lines in the figures. The RMSE can be used as an uncertainty in recovered estimates of the hindcast parameter deviations (also shown in Fig. 2, as blue points with error bars on the vertical axes; discussed later, in Sect. 4).

The performance of the neural network model, averaged over the four possible training/test splits, is summarised in Fig. 3. Figure 3a shows, by region, the number of parameters that can be recovered with a reasonable amount of skill, defined here as $R^2 > 0.3$, which is usually sufficient to at least give information on the sign of the parameter perturbation. Most tropical and subtropical regions, along with the Antarctic, show 7–10 parameters recovered adequately, while only 2–5 parameters can be recovered in most midlatitude regions. In part this likely points to the value of the additional subsurface ocean drifts, which are only available in tropical regions, and which show relatively consistent drifts over the first few months. Ocean parameters are therefore typically recovered most accurately, with ISOPYC, VDIFFSURF, MLLAM and MLDEL (see Table 2) all exceeding $R^2 = 0.3$ in at least 25 of the 65 regions. The less effective parameter recoveries in midlatitudes may occur also because large-scale dynamical fields, varying strongly on synoptic timescales, can exert a more non-linear impact on model drifts than is the case at lower latitudes, diminishing the ability to detect any local parametric control.

Figure 3b shows the largest $R^2$ found for each parameter, and the region in which this occurs. In total, there are 15 parameters with maximum $R^2$ values greater than 0.5, and 20 with $R^2 > 0.3$, although some significant skill (above the 99% level, via a $t$ test) can be found for all parameters except those associated with atmospheric chemistry. The ranking of parameters according to $R^2$ highlights those which have the biggest impact on model drift in the first month, based on the monitored variables, which are weighted in favour of the near-surface. Some parameter perturbations, particularly those relating to upper ocean processes (coloured blue), can be recovered well, especially in tropical regions (where subsurface ocean fields are available). Two parameters relating to sea ice (ALPHAM and DTICE) can also be recovered well, but only in the Antarctic region, where (in December) the melt season is underway. Atmospheric chemistry parameters (coloured green) cannot be meaningfully recovered in any region, indicating that they have almost no impact on the variables being monitored on this timescale. The range of regions included in Fig. 3b shows the extent of the regional variability in the impact on drifts of each parameter.

The dashed and dotted lines in Fig. 3b show the best $R^2$ values for parameter recovery using month-2 and month-6 drifts, respectively. It can be seen that the accuracy of the perturbation estimates, in most cases, decreases from month 1 to month 2, and decreases further by month 6, showing that model drifts become less separable as the simulations progress, due to non-linear interactions and non-local influences from outside the region. Drifts are distinct, at least in the ocean, as month-1 means, but a greater separability could be achieved by using shorter lead times.

Also shown in Fig. 3b (as thin bars) are the $R^2$ values for parameter recoveries using global-mean drift fields. Recoveries are generally less accurate than the best regional recovery, as expected, although results differ greatly among parameters. For the two sea ice and several of the ocean mixing parameters, the global recovery has almost no skill, implying that the global-mean model drifts contain very little signal from perturbations to these parameters. This suggests that tuning parameters using global mean drift fields may often not be useful, and highlights the value of modelling regional drifts. Note that we do not attempt any global recoveries in Sect. 4, either using global-mean drifts or combining multiple regional drifts, as we do not expect process errors to present uniformly across different regions in INIT.

Results from similar evaluations of PP_1980 and PP_IC, as sensitivity tests with respect to ensemble size and initial state variability, respectively, are given in Appendix "Recovery sensitivity tests". Drifts in some atmospheric variables were found to vary substantially with start date, but it was found still to be beneficial to include these drifts in the neural network, even with single start dates; the use of multiple start dates further reduces this problem of state dependence.

### 3.3 Dominant parameters and processes causing model drift

The accuracy with which unknown parameter perturbations can be estimated, as measured by $R^2$, can be used to rank the 'most important' processes in each region. Figure 4 shows the best-recovered parameter in each region, which is equivalent to the parameter that has the largest impact on PP_FULL drifts in the first month, averaged across all the atmospheric and near-surface ocean fields (subsurface ocean fields are excluded here to ensure consistency between tropical and extratropical regions). The atmospheric entrainment coefficient (ENTCOEF) has the biggest impact in most tropical regions, causing substantial drifts in several fields, while the wind mixing energy parameters (MLLAM and MLDEL) are dominant in the midlatitudes, particularly over the Southern Ocean. The sea ice albedo (ALPHAM) is most important in the Antarctic region, but ocean mixing parameters (ISOPYC, VDIFFSURF) are more important in Arctic regions. Note that drifts are calculated in boreal winter (December), and some results may be different at other times of year, for example if Arctic sea ice drifts are larger or have greater impact in boreal summer, when ice cover is more marginal. The $R^2$ values in the plot again show the regional variation in accuracy with which any parameter information can be recovered.

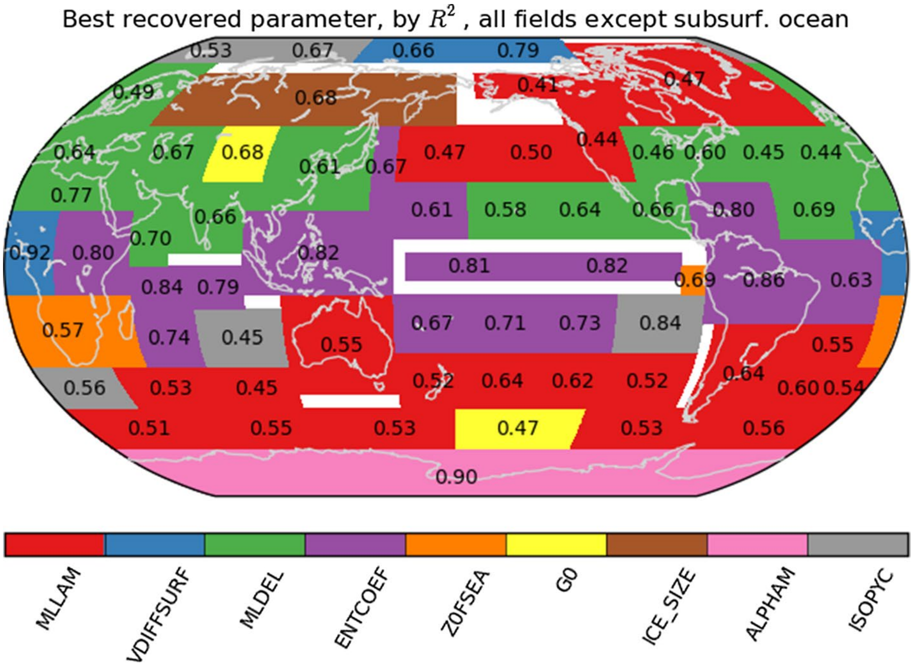

**Fig. 4** Map showing the parameter that is best recovered in each region, as measured by $R^2$, using a model similar to that tested in Figs. 2 and 3 but excluding drifts in the subsurface ocean $T$ and $S$ fields. The relevant $R^2$ value is printed in each region. *White areas* are not covered by any of the regions (see Fig. 3a)

Several other metrics for ranking the parameters according to their impact on drifts were explored (not shown), and these showed broad agreement in the parameters that rank highly, particularly regarding the dominance of the atmospheric entrainment coefficient (ENTCOEF) in tropical regions. These ranking methods provide a way of identifying the parameters and, by extension, the physical processes, to which short term model drifts are most sensitive. Overall, the results highlight the importance of convective processes on sub-seasonal timescales (see also, Ma et al. 2014), where the main sources of predictability are in the tropics.

## 4 Diagnosis of hindcast process errors

### 4.1 Standard HadCM3 hindcasts

The neural network was then applied to the set of 8 initialised hindcasts, to identify which physical processes, linked to parameter values, are most likely to be in error in each region, as detected through the hindcast ensemble drifts. Drifts in the same set of coupled model fields as above were calculated for the first month from each of the 8 start dates, and these were used as inputs to the neural network. Parameter perturbations were estimated separately for each start date, and averaged. The tests performed in Sect. 3 were used to recover parameters only where a reasonable amount of confidence may be achieved; that is, all region-parameter combinations where recovery accuracy from PP_FULL has $R^2 > 0.3$. The results can be understood as the deviation of each parameter from the value that would minimise the overall drift in the first month (i.e., minus the change that should be applied to the control parameter value in order to minimise the drift), in each region. More generally we interpret these deviations as proxies for physical process biases in the forecast model, which may also have a structural component.

Figure 5 shows the results of the parameter recoveries, with the additional strong constraint that the magnitude of the recovered deviation must be larger than the recovery RMSE (Fig. 2) for that region and parameter in the perturbed physics set, such that deviations can be identified as non-zero. The RMSEs calculated in Sect. 3 are likely pessimistic estimates, since the test set consisted of drifts from a single year, while recoveries for the hindcasts are made separately using 8 different sets of drifts, and then averaged. Therefore, Fig. 5 shows a conservative estimate of the necessary parameter adjustments that can be robustly determined from the INIT hindcasts. As an example of the interpretation of this figure, the Southern South America (SSA) region shows significant positive perturbations in parameters MLLAM, EACF, and G0, requiring downward adjustments to reduce drifts, and a significant negative perturbation in ENTCOEF, necessitating an upward adjustment. From this it is suggested that wind-driven mixing, adjusted cloud fraction and boundary layer stability may be too large in this region, while atmospheric convection may be extending too high (entrainment rate ENTCOEF is too low), in the standard HadCM3 model.

Regionally varying process perturbations are expected in this case because model biases are likely to be different

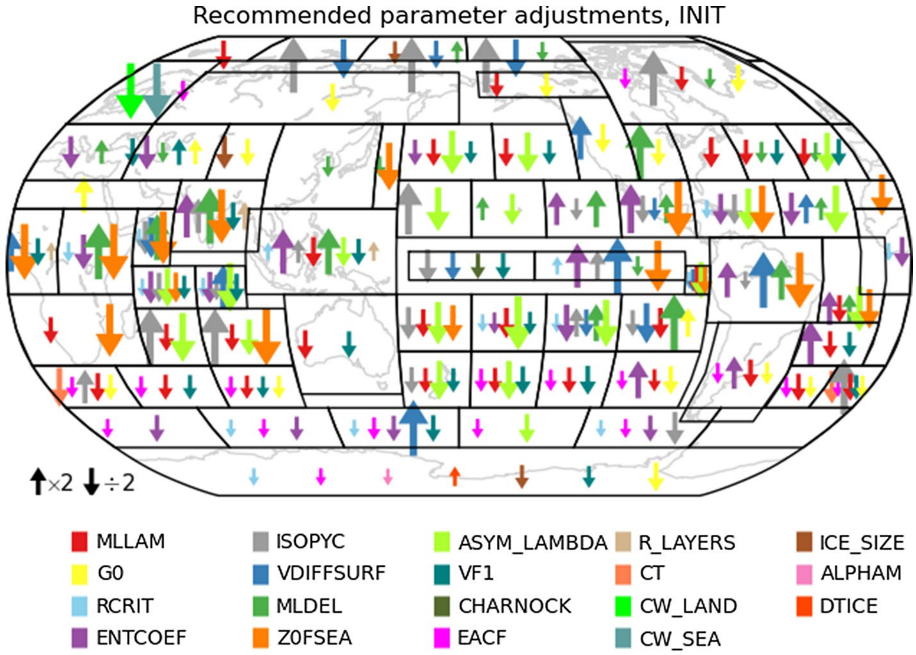

**Fig. 5** Recommended parameter adjustments for all significant deviations detected in $\Delta p$, by region. An *upward* (*downward*) *arrow* marks a negative (positive) $\Delta p$, which would be corrected by adjusting the parameter in the positive (negative) direction. *Arrows* are sized by the multiplicative change to the control parameter required by the $\Delta p$, so that a doubling in value ($\Delta p = +1$) is comparable to a halving ($\Delta p = -0.5$), although *arrow* length saturates at $-0.75$ and $+3$ for readability ($\Delta p$ values cover a range $[-0.9, +5.0]$)

in different regions, depending on the processes that dominate local weather and climate. This is the reason that the parameter recovery is run region-by-region, rather than combining drift information from multiple regions. Nevertheless, it is encouraging to see substantial coherence between neighbouring regions, as would be expected where climate regimes extend over larger areas. Broad classes of regionally distinct bias behaviour can be identified; e.g., in the tropical Indian Ocean, over land in Asia and North Africa, or in the southern Atlantic and over South America. The regional differences suggest that spatially heterogeneous parameter values might be needed to minimise hindcast drifts.

The most consistently recovered deviations are in the ocean wind mixing energy scaling factor (MLLAM), the atmospheric convective entrainment coefficient (ENTCOEF), and the ocean isopycnal diffusion factor (ISOPYC), consistent with Figs. 3 and 4. Drifts suggest that wind-driven mixing is almost exclusively too strong, while convective entrainment and isopycnal mixing appear to be more regionally varying, with roughly equal numbers of regions with increases and decreases to the parameters. For example, isopycnal mixing appears to be substantially too weak in several southern Atlantic, Indian and Arctic regions, but too strong in parts of the southern Pacific and around Central America.

The atmospheric boundary layer stability parameter (G0) is found to be too large (meaning mixing of heat and momentum are too weak) in several regions in the Southern Ocean, and over northern America and north-eastern Asia, but too small in western Asia and northern Africa. The ocean mixing decay rate with depth (MLDEL) is found to be too small in parts of the tropical Indian and eastern sub-tropical Pacific Oceans. In the Antarctic, where sea ice albedo parameters, ALPHAM and DTICE, can be estimated with relatively high accuracy (Fig. 3b), small downward and upward adjustments, respectively, are suggested.

The sea surface roughness length (Z0FSEA) requires particularly large reductions in the tropics, suggesting that model atmosphere-ocean coupling, in terms of heat and momentum fluxes, is too strong. Note, however, that the range of (normalised) Z0FSEA values sampled in PP_FULL is larger than for many other parameters (Table 2). This suggests that the value of Z0FSEA is poorly understood, or that it varies regionally and therefore may not be well captured by a single control value, or both.

### 4.2 Modified HadCM3 hindcasts

To test the validity of the suggested parameter changes derived above, the initialised hindcast set was rerun (INIT_MOD) with a globally uniform adjustment made to one parameter, the wind mixing energy scaling parameter, MLLAM. As can be seen from Fig. 5, MLLAM is suggested to be too large in 29 different regions, and is not found to be significantly too small in any regions, so a downward adjustment to this parameter should lead to reduced drifts across a large global area. Typical recovered $\Delta p$ in the relevant regions were ~0.5–1.0, suggesting reductions of ~33–50%, so a conservative reduction of around 30% was made, replacing the standard MLLAM value of 0.7 with a new value of 0.5.

Normalised drifts in the 29 target regions (from Fig. 5) were indeed reduced in INIT_MOD relative to INIT, as

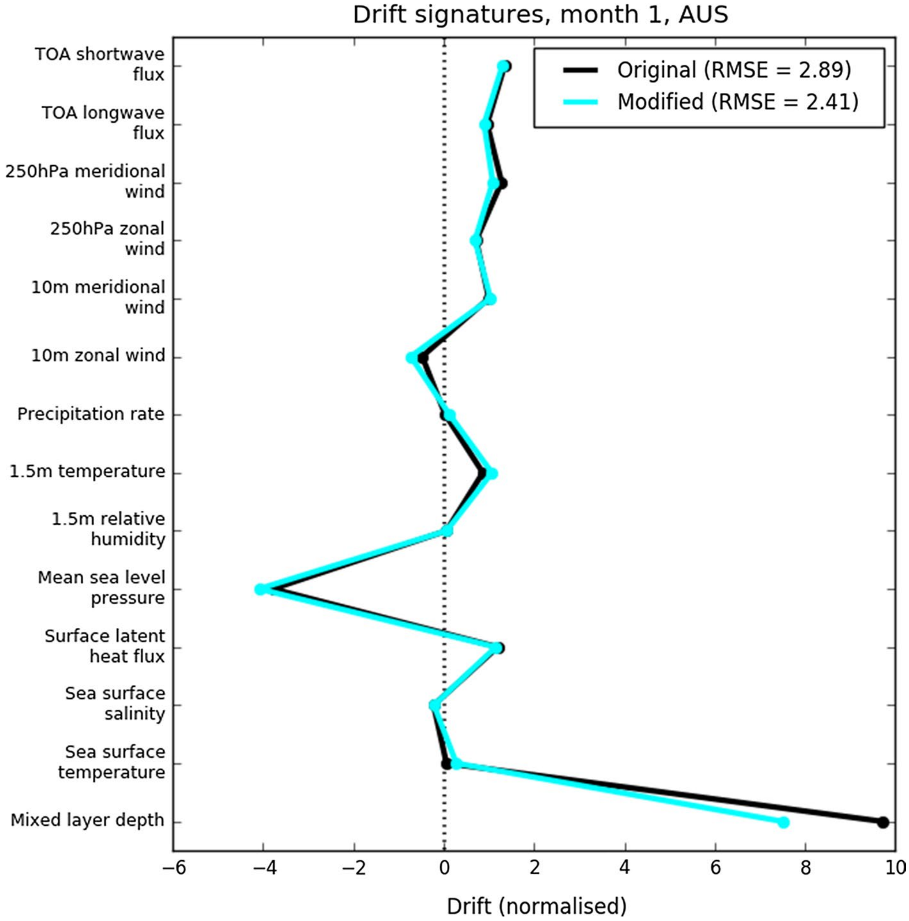

**Fig. 6** Example drift profiles for the Australia (AUS) region from INIT (*black*) and INIT_MOD (*cyan*)

measured by root-mean-square amplitude of the drifts across all model fields. An example of this change in the drift signature is shown in Fig. 6, for the Australia (AUS) region. In this case, and in several other regions, the RMSE is dominated by errors in the ocean mixed layer depth, which are particularly large when normalised by the reference interannual variability, and the reduction in RMSE occurs primarily via this field. Across all fields, the normalised RMSE in month 1 relative to the standard reference is reduced by around 15%. The recommendation of a downward adjustment to MLLAM was specific to this choice of drift fields and normalisation method, and reduced drifts in other model fields could, in principle, be targeted by altering the method to give more weight to those fields.

By linearly regressing RMSE against the 31 $\Delta p$ values using PP_FULL, it is possible to predict the approximate changes in RMSE that would be expected in INIT_MOD following the reduction of MLLAM to 0.5, by using the regression coefficient for the MLLAM perturbation. Figure 7 plots the actual change in RMSE against the expected change, for all 65 regions. Regions in which large decreases in MLLAM were recommended (filled circles) generally do

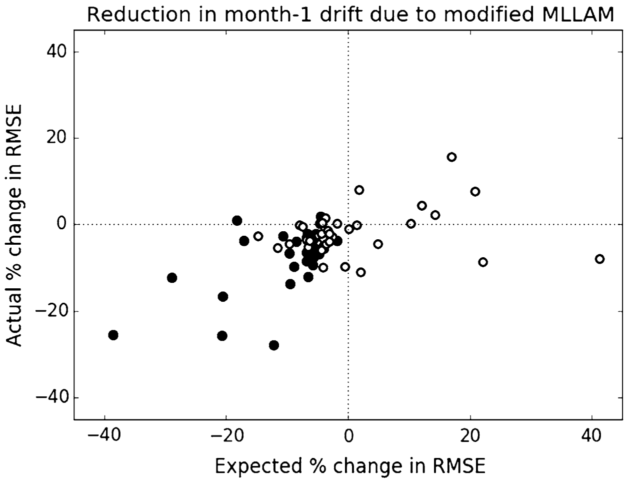

**Fig. 7** Percentage change in month-1 drift from INIT to INIT_MOD, measured as the RMSE of all monitored fields, plotted against the change predicted from PP_FULL for a 29% reduction in MLLAM. The 29 target regions in which a reduction in MLLAM was robustly recommended are plotted as *filled circles*, and regions in which recommended changes were not statistically significant are plotted as *open circles*

show the largest reductions in RMSE. The mean reduction over these 29 target regions is 8%, and improvements are seen individually in 26 of these 29 regions. Most of the other regions, where recommendations are less robust, also show small reductions in RMSE, and the global mean change is a reduction of around 5%, suggesting that the global control MLLAM value may be too large. However, in several of the Southern Ocean and Antarctic regions, the control value of MLLAM is suggested to be too low (positive changes recommended), and when MLLAM is reduced in INIT_MOD, the RMSE in most of these regions does increase, as the drift is exacerbated (see the upper right of the figure).

The correlation between the expected and actual changes in RMSE in INIT_MOD (0.50, using all 65 regions; $p < 0.001$) strengthens confidence that the recovery method translates well from the perturbed physics ensemble to the initialised hindcasts, and that the $\Delta$p recoveries from INIT are an accurate guide to changes in parameters that reduce drifts, even beyond the conservative selection shown in Fig. 5. It can be expected that further reductions in RMSE to those shown in Fig. 7 could be achieved by altering other parameters following Fig. 5, and, furthermore, by making the adjustments on a regional rather than a global basis. The latter would be particularly beneficial for parameters such as VDIFFSURF or ENTCOEF, where the suggested adjustments show substantial regional variability.

## 5 Discussion

The results presented above suggest that various parameters in the standard control version of HadCM3 are sub-optimal for minimising regional drifts in several model fields, when hindcasts are initialised with the full-field DePreSys atmosphere and ocean analyses. The set of fields used to measure drifts is weighted towards the near-surface, where reducing drifts is of particular importance, since errors in the lower atmosphere and upper ocean can become amplified over time through coupled feedbacks. Large regional variability was seen in the suggested adjustments for several of the parameters, indicating that the control values may be realistic in a globally averaged sense, but that regional model drifts could be reduced by modifying the parameters regionally. Parameters ENTCOEF, ISOPYC, MLLAM, MLDEL, VDIFFSURF and G0 were all confidently determined to deviate substantially from their optimal values in many regions, pointing to their importance to forecast drift on monthly timescales.

In all cases the diagnosed parameter deviations were within the range explored in the perturbed physics experiment PP_FULL, although often near the limit of this range (Fig. 2). This is perhaps not surprising since the PP_FULL range was originally selected to give long-term climatically reasonable solutions using only global parameter adjustment, and the neural network is limited to this training domain. Indicated perturbations should therefore be taken as conservative estimates of changes which could be made to region-specific parameter values in order to minimise month-1 drifts. As examples, for two of the parameters featured in Fig. 5 (EACF and R_LAYERS), making the recommended adjustment would require a shift beyond the range included in PP_FULL (Table 2), presumably because the range is overly restrictive in the context of regional parameter settings.

Hindcast drifts may also point to structural model errors which will be less susceptible to reduction by tuning the parameter values (Severijns and Hazeleger 2005). In PP_FULL, the only errors present are those caused by the parameter perturbations, while in the initialised hindcasts additional structural or systematic errors will be present. Perhaps the relationships between parameter perturbations and drifts established for PP_FULL may, therefore, not transfer cleanly to INIT? The success of the test described in Sect. 4.2, however, shows that there does exist correlation between parameter-drift relationships in the absence of and in the presence of structural or other systematic errors, and therefore that a neural network trained on perturbed physics ensemble data can provide useful insights into modelling initial drifts in a real-world trajectory using regional parameter variations. Since model parameters are closely associated with physical and dynamical processes, it makes sense that even systematic-error-induced drifts may be partly explained by parameter changes.

One could imagine using a direct method to reduce month-1 drifts by directly perturbing parameters in the initialised hindcasts themselves and seeking to reduce drifts iteratively. Such a method would be similar to the use of variational data assimilation for parameter estimation problems (e.g., Eknes and Evensen 1997), in which a cost function (namely month-1 drifts) is minimised through iterative parameter adjustment. There is no necessary assumption in applying such methods that the model is structurally correct; the success is measured by reductions in the cost function alone. Other tests might be developed to investigate the representation of structural drifts through parameter tuning. Parametrisations often have alternative representations in models, and twin experiments might be performed tuning one set of parameters against model data developed with structurally different parametrisation schemes.

Another potential question over the usefulness of this method is whether minimising drift is necessarily beneficial to forecasts. It is common, for example, to correct for drift errors in forecast post-processing, by calculating the mean error signal from a set of calibration hindcasts. However, any non-linear effects occurring during the forecast cannot be removed via post-processing. A good example is the occurrence of atmospheric convection in the tropics, which is controlled in part by parameter ENTCOEF. The initiation is highly non-linear and very sensitive to SST thresholds. Therefore, particularly in the tropical regions, which offer the largest potential predictability on the sub-seasonal timescale, it can be expected that minimising forecast drift will lead to a more skilful forecast.

The possibility of making regionally varying adjustments to parameters deserves further discussion. While an ideal parametrisation should, arguably, be completely physically based, with regional variation arising only through the interaction of the homogeneous parameters with the local environment, most parametrisations fall somewhat short of this ideal and instead rely strongly on empirical formulation. For example, parametrised ocean isopycnal mixing may need to vary in strength regionally, dependent on the unresolved eddies present in various ocean circulation regimes (Visbeck et al. 1997; Forget et al. 2015). It is therefore realistic and, in principle, numerically feasible for any of the parameters discussed above, such as atmospheric entrainment rate, wind mixing strength, or tracer isopycnal diffusion rate, to assume regionally varying values, provided this is done in a smoothed manner to avoid sharp parameter changes at regional boundaries. Indeed, regional parameter adjustments have previously been undertaken as part of data assimilation frameworks (e.g., Annan et al. 2005; Zhang 2011). The results of Sect. 4 therefore strongly support the case for spatial variation in parameter values, as an empirical means of reducing forecast drift.

Indeed, extending the arguments of Rodwell and Palmer (2007) and Ma et al. (2014), a climate model tuned with regional process parameters derived from short term drift metrics, as suggested here, may also perform better with regard to longer-timescale climate modelling. The indication here, that many parameters require adjustments in different directions in different regions, suggests that standard versions of climate models, with well-tuned but globally constant parameters, can fall prey to regional drifts. These drifts are bound to interact non-linearly and in complex ways across regions, in long-term simulations, which complicates efforts to improve climate model performance through a global parameter tuning approach, while considering processes more locally may be more successful. Such regional parameter tuning may or may not be appropriate when using the model for less well-constrained applications such as future climate projection.

Several aspects of these experiments were strongly constrained by the capability of existing systems. DePreSys initial conditions for the hindcasts were only available for 1 November each year, while the CPDN version of HadCM3 used for the perturbed physics experiments could only be run from 1 December, leading to a slight mismatch, to the extent that model drifts exhibit seasonal variation between November and December. In addition, our choice of month 1 as the lead time for drift comparisons was constrained by the output frequency available from the CPDN system, which is designed primarily for longer-term climate simulations. Both hindcast and perturbed physics drifts could usefully be studied on much shorter timescales, at least down to a few days, when this method would bear comparison to the 'Transpose-AMIP' (Transpose Atmospheric Model Intercomparison Project II; Williams et al. 2013) method, in which climate models are initialised using operational numerical weather prediction analyses to study short-term drifts. At such lead times, drifts in atmospheric fields should be more separable and therefore of greater value in estimating parameter errors, which may permit the recovery of information on a wider range of parameters. The use of short model runs to understand climate model biases is an increasingly active direction of research (e.g., Klocke and Rodwell 2014; Wan et al. 2014), and the perturbed physics ensemble approach is a useful addition to this set of techniques.

## 6 Summary

These results illustrate a new use of the perturbed parameter modelling framework by focusing on transient model behaviour under known physical perturbations, which has not to our knowledge been previously studied, and using this information to analyse initialised hindcast drifts. A number of perturbed physics ensemble runs on the climate*prediction*.net distributed computing system have been used to interpret the transient drifts (errors) that are produced in HadCM3 coupled atmosphere-ocean forecasts, initialised with the full-field DePreSys analysis. In the perturbed physics framework it has been shown that model drifts in the first month can be represented as a combination of different model process errors caused by deviations in physical parameter values. As a result, groups of regionally averaged drifts across different model variables can be used to infer parameter perturbations that would give rise to these drifts. This technique highlights both those parameters which exert a strong influence on near-surface forecast fields (both ocean and atmosphere), and those parameters which are largely irrelevant to one-month forecasts (e.g., those associated with atmospheric chemistry).

The perturbed physics ensemble was used to train a regional set of neural networks which predict parameter perturbations based on the transient drifts. This system was then used to infer equivalent parameter perturbations indicating the physical processes that may be causing regional drift in initialised HadCM3 hindcasts. This then indicates the regional parameter value changes that have the potential to reduce drifts by reducing the relevant regional process errors. Upper ocean parameters were found to be particularly influential, along with the atmospheric convective entrainment coefficient. A range of regional adjustments, requiring both strengthening and weakening of many processes, were recommended for these parameters. The method was tested by altering, at the global level, the wind mixing energy scaling factor (MLLAM) in a new set of initialised hindcasts, leading to reductions in month-1 drifts in the regions previously identified as sensitive in the standard HadCM3 model.

It is suggested that such a method could be a practical means of improving the skill of short-timescale initialised coupled model forecasts, and by extension of understanding and reducing climate model biases. The possibility of developing better, regionally tuned, long-range climate model simulations using these methods is especially inviting. From a computational point of view, the ensemble size of 2000 members used here is relatively modest, and could potentially be reduced further if the perturbed parameter set were restricted to a smaller number of more influential (on short timescales) parameters. In any case, the short-range simulations needed would not require huge resources, even for state-of-the-art, high-resolution climate models. The method may therefore be a viable approach to empirically tuning climate model parameters on a regional basis to target reduced drift at various forecast lead times, as well as potentially for long-range climate runs. Further work is now underway to modify HadCM3 to allow for regional parameter adjustment in order to test some of these ideas.

**Acknowledgements** We thank two anonymous reviewers for their helpful and constructive comments, which have improved this manuscript. We are very grateful to Simon Wilson at the National Centre for Atmospheric Science's Computational Modelling Services, for his work to make available a version of HadCM3 with which it was possible to run the initialised hindcasts. We thank the Met Office Hadley Centre for their technical and scientific support for the development and application of CPDN. We also thank Doug Smith for supplying the restart files for the DePreSys hindcasts, Jon Robson and David Schroeder for help with accessing the DePreSys reference data, and the CPDN participants for running the perturbed physics ensembles. The work was funded by the UK Natural Environment Research Council (NERC) (project ERGODICS). K. Haines is partly supported by NERC under the National Centre for Earth Observation. Perturbed physics and initialised hindcast drift data are available on request.

## Appendix

### Statistical model of parameter-drift relationships

The prediction of parameter perturbations given drift inputs was performed using a neural network, with three units in the hidden layer, specifically using the *nnet* package in R. A separate neural network was trained for each combination of region and parameter, with the set of regionally averaged month-1 drifts as inputs to the first layer and the parameter perturbation $\Delta$p as the output, the mean-squared-error in which was minimised. These neural networks were then used to estimate unknown parameter perturbations, using drift field inputs. Several other machine learning algorithms, including multivariate linear regression and multi-layer perceptron, were tested, and the neural network was found to perform best. Differences from other methods were small, however, suggesting that drift interaction terms, which can be modelled by the neural network but which were not included in the linear regression, provide a limited amount of information at this lead time. The performance of the neural network was not formally tuned, since the aim was to demonstrate that the method of inferring parameter deviations from forecast drifts is practically useful, which requires only a certain level of skill in the parameter recovery process to be measured. Tuning of the recovery algorithm could improve, to a small degree, the performance of the method, and therefore would be recommended were this method to be used in practice.

Examples of the output of the neural network model, further to those shown in Fig. 2, are given in Fig. 8. Note that in regions that are completely insensitive to a parameter, the neural network can return one of a limited set of predicted values, as for DTICE in non-polar latitudes, but these can be safely ignored.

### Recovery sensitivity tests

The parameter recovery method was first tested using drifts from a single start date, with the same date for both training and test sets. These sets were formed by splitting the large, 2200-member set PP_1980 into groups of 2000 and 200 members respectively. Neural networks were trained on drifts and parameter perturbations from the training set, and then used drifts from the test set to predict the parameter perturbations responsible. The results of this procedure are shown in Fig. 9a (wide bars), which can be compared to Fig. 3b. The accuracy of the parameter estimations is greater than in Fig. 3b, due to the greater consistency between training and test member drifts produced for the same start date (i.e., with the same start date, two members

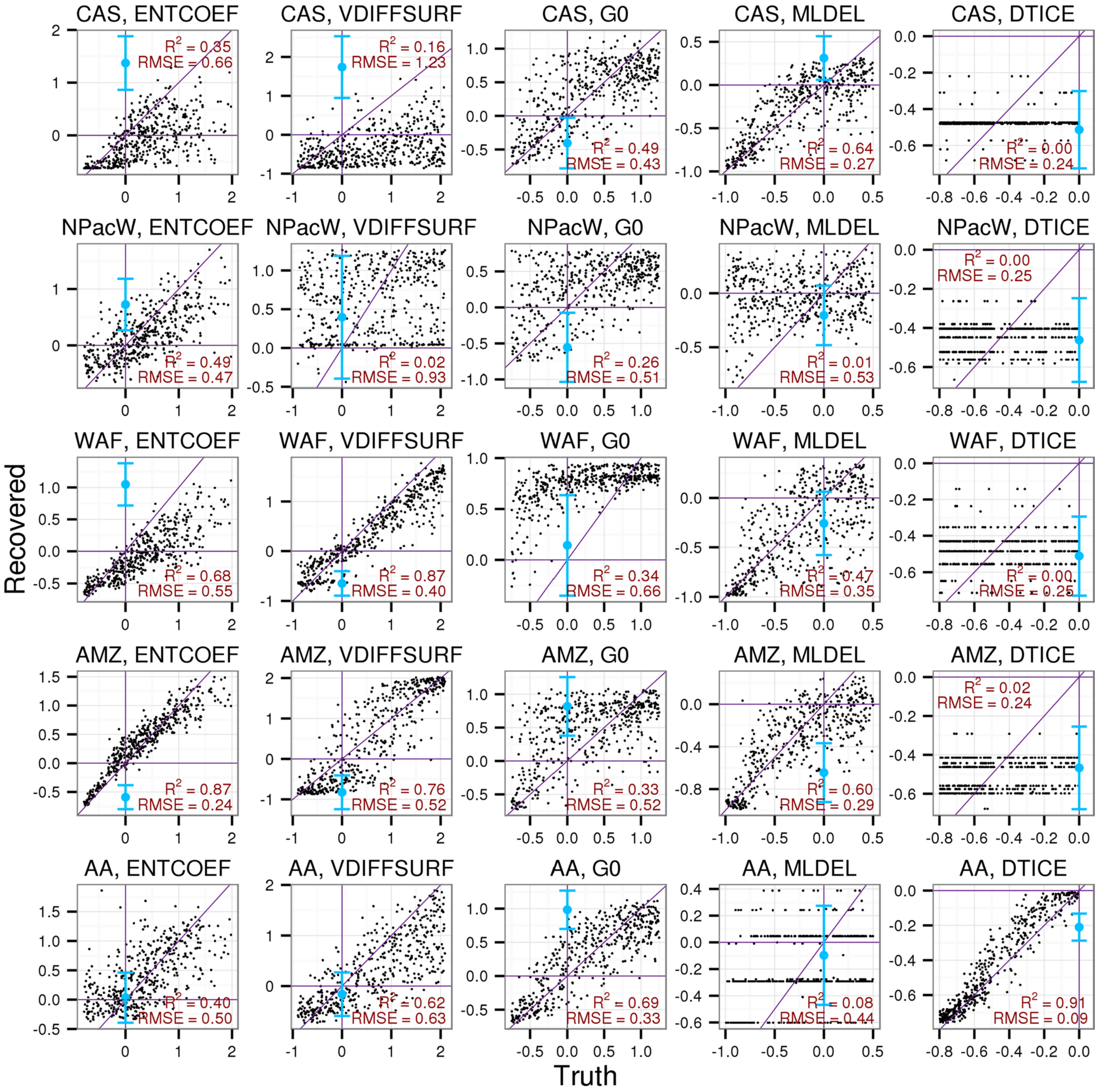

**Fig. 8** Examples of parameter recovery for several regions and parameters: predicted $\Delta$p values are scattered against the true values, for the 500 '1960' members of PP_FULL, which are used as the test set for these examples (the training set consisted of the other 3 years: '1940', '1950', and '1980'). Parameters shown are (*left* to *right*) ENTCOEF, VDIFFSURF, G0, MLDEL, and DTICE, in regions (*top to bottom*) CAS (central Asia), NPacW (northwestern Pacific), WAF (western Africa), AMZ (Amazon basin), and AA (Antarctic). The $R^2$ and RMSE values for the group of recoveries are printed in each plot. The *blue points* show the value predicted for INIT (see Sect. 4), with the true RMSE (calculated as the average of the four possible training/test splits) shown as the *error bar* on this estimate. See Fig. 3a for region definitions

with similar $\Delta$p values are very likely to produce similar drifts, whereas with different start dates this is not necessarily the case).

The degradation in parameter recovery accuracy when the drifts are taken from a different start date to the training set (as is the case if the method is applied to initialised hindcasts) is shown by the thin bars in Fig. 9a. For this calculation, the test set of PP_1980 members was replaced by a set of 50-member ensembles from each of 6 different start dates (that is, 300 of the 350 members in

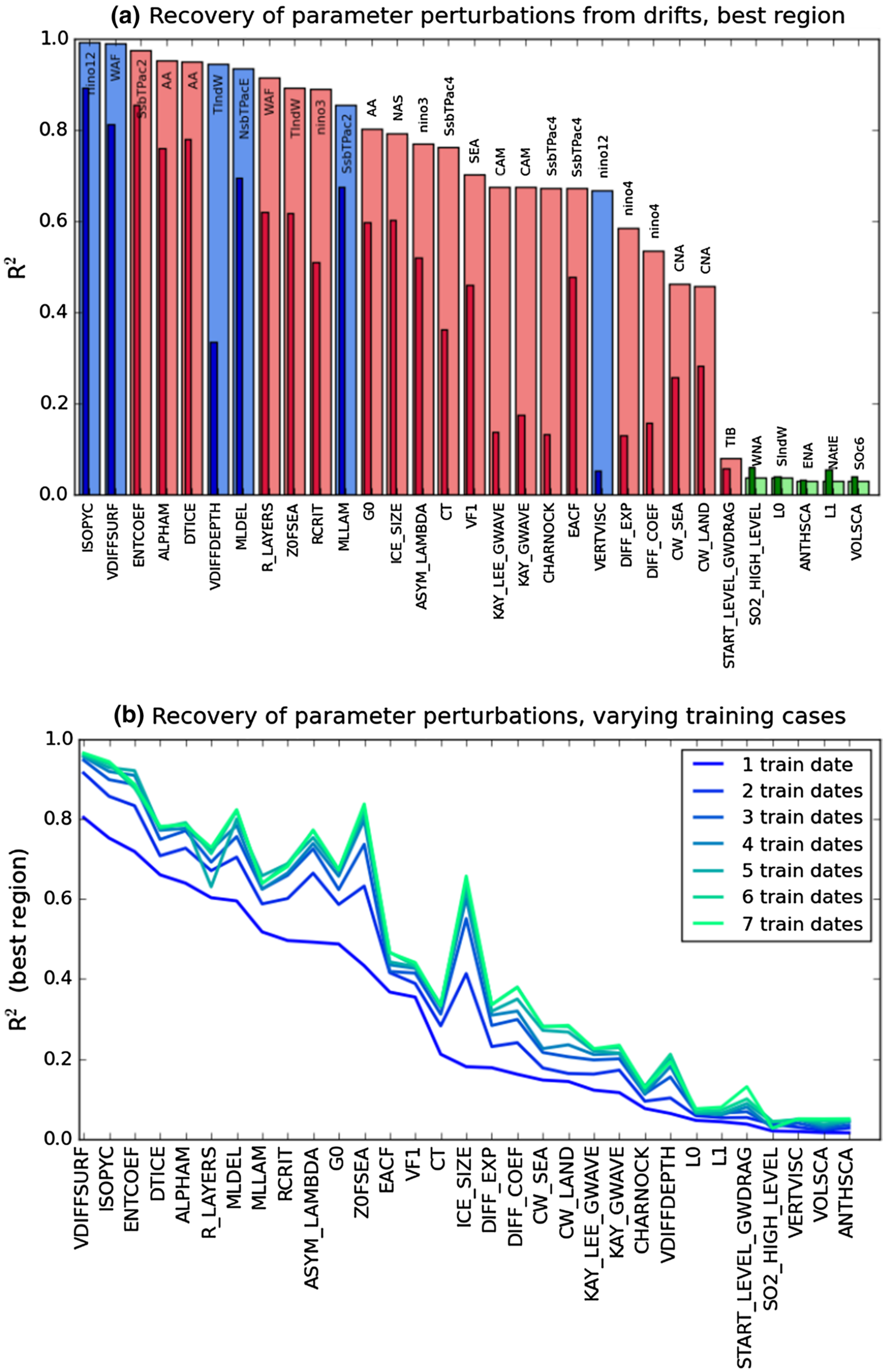

**Fig. 9** **a** Verification of the neural network parameter recovery method on CPDN data, showing the region with highest R$^2$ for each parameter, *coloured* by parameter type (physical atmosphere (including sea ice) in *red*, ocean in *blue*, chemistry in *green*); using the same initial conditions (‘1980’) for the training and test sets (*wide*, *pale bars*, with the region producing the highest $R^2$ printed above each *bar*), and different initial conditions (‘1930’–‘1980’) for the training and test sets (*thin*, *dark bars*). **b** As in **a**, but for training sets of 40 members from $n_{\text{train}}$ starting conditions, and test sets of 70 members from all 7 starting conditions with predictions averaged over the start years. In **b**, parameters are ordered from*left* to *right* according to the $R^2$ values for $n_{\text{train}} = 1$, and *lines* are *coloured* by $n_{\text{train}}$ value, from *blue* (1) to *green* (7)

PP_IC, excluding the 50 members beginning in ‘1980’). These 6 initial states were taken from ‘1930’ to ‘1990’, sampled every 10 years (excluding ‘1980’), and spanned the range of model ENSO states (Niño3.4 SST values from −0.9 to 2.0 K). The figure shows that $R^2$ values are reduced compared to the ‘1980’-only case, but that the model still shows significant skill, as 17 of the parameters can be estimated with $R^2 > 0.3$. Ocean parameters are still the most accurately recovered, and the ordering of the parameters by $R^2$ has not greatly changed. A notable exception to this is the background vertical viscosity (VERTVISC), which has fallen from an $R^2$ of around 0.7 to near-zero, indicating that the effects of changes to this parameter are seen most strongly in drifts in fields that vary greatly between start dates.

To mitigate against this sensitivity, drifts from several different start dates may be combined to give a larger training set that is less strongly linked to one particular start date. Similarly, predictions of parameter perturbations can be made separately for each start date, and then averaged. To measure the benefit resulting from this approach, the set of seven 50-member ensembles of PP_IC were used, and drifts from a number of start dates were included in the training set.

For each parameter and region, and for each number of training start dates used, $n_{\mathrm{train}}$, the seven ensembles were split into 40-member training sets and 10-member test sets. $n_{\mathrm{train}}$ of the training sets were chosen at random and their drifts used in combination to train a neural network. The $\Delta$p were predicted separately for all seven start dates and then averaged for each set of parameter perturbations, giving 10 $\Delta$p predictions with which to compute $R$ for the model. This process was repeated 1000 times (using different training/test splits) and the results were averaged, to ensure that favourable combinations of training and test years did not skew the results for small $n_{\mathrm{train}}$, and because of the large variability in $R$ resulting from the small number of test cases. Finally, for each parameter, the largest $R^2$ among all the regions was used as the measure of recovery accuracy. Note that $R^2$ values are likely to be artificially high, compared to a practical situation involving initialised hindcasts, due to the presence of the same start year(s) in both training and test sets, but the sensitivity to $n_{\mathrm{train}}$ is the focus of this calculation, rather than absolute $R^2$ accuracy.

Figure 9b shows that, at least with the small ensembles used in this sensitivity test, there is benefit to combining drifts from more than one start date. For those parameters which are best recovered, the improvement is seen primarily increasing from one to 3–4 start dates, and further increases offer relatively little benefit. With this in mind, four start dates were chosen for PP_FULL, as introduced in Sect. 2.

We also tried an alternate order of operations, with drifts averaged across the different start dates before being used in the training and test sets, resulting in a single set of $\Delta$p predictions. Compared to making individual recoveries for each date and averaging the $\Delta$p, this method was found to perform slightly worse, so was not used. The reason for this may be that the neural network is able to make use of more information when combining training drifts from multiple start dates than when the drifts are averaged across the start dates.

Finally, for the training ensembles, tests using the 2000-member '1980' ensemble (not shown), again with $R^2$ as the metric, showed that similar results could be obtained with a reduced ensemble size of around 500 members. This was therefore the size chosen for the four start date ensembles in the PP_FULL set.

## References

Alessandri A, Borrelli A, Masina S, Cherchi A, Gualdi S, Navarra A, Di Pietro P, Carril AF (2010) The INGV–CMCC seasonal prediction system: improved ocean initial conditions. Mon Weather Rev 138(7):2930–2952

Allen MR (2003) Climate forecasting: possible or probable? Nature 425(6955):242–242

Annan JD, Hargreaves JC, Edwards NR, Marsh R (2005) Parameter estimation in an intermediate complexity earth system model using an ensemble Kalman filter. Ocean Model 8(1):135–154

Brierley CM, Collins M, Thorpe AJ (2010) The impact of perturbations to ocean-model parameters on climate and climate change in a coupled model. Clim Dyn 34(2–3):325–343

Collins M, Booth BBB, Bhaskaran B, Harris GR, Murphy JM, Sexton DMH, Webb MJ (2011) Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model ensembles. Clim Dyn 36(9–10):1737–1766

Ding H, Greatbatch RJ, Latif M, Park W (2015) The impact of sea surface temperature bias on equatorial Atlantic interannual variability in partially coupled model experiments. Geophys Res Lett 42(13):5540–5546

Doblas-Reyes FJ, García-Serrano J, Lienert F, Biescas AP, Rodrigues LR (2013) Seasonal climate predictability and forecasting: status and prospects. Wiley Interdiscip Rev Clim Change 4(4):245–268

Eknes M, Evensen G (1997) Parameter estimation solving a weak constraint variational formulation for an Ekman model. J Geophys Res 102(C6):12479–12491

Forget G, Ferreira D, Liang X (2015) On the observability of turbulent transport rates by Argo: supporting evidence from an inversion experiment. Ocean Sci 11(5):839

Giorgi F, Francisco R (2000) Uncertainties in regional climate change prediction: a regional analysis of ensemble simulations with the HADCM2 coupled AOGCM. Clim Dyn 16(2–3):169–182

Gordon C, Cooper C, Senior CA, Banks H, Gregory JM, Johns TC, Mitchell JF, Wood RA (2000) The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. Clim Dyn 16(2–3):147–168

Jung T (2005) Systematic errors of the atmospheric circulation in the ECMWF forecasting system. Q J R Meteorol Soc 131(607):1045–1073

Klocke D, Rodwell M (2014) A comparison of two numerical weather prediction methods for diagnosing fast-physics errors in climate models. Q J R Meteorol Soc 140(679):517–524

Knight CG, Knight SHE, Massey N, Aina T, Christensen C, Frame DJ, Kettleborough JA, Martin A, Pascoe S, Sanderson B, Stainforth DA, Allen MR (2007) Association of parameter, software, and hardware variation with large-scale behavior across 57,000 climate models. Proc Nat Acad Sci 104(30):12259–12264. doi:10.1073/pnas.0608144104

Kumar A, Chen M, Zhang L, Wang W, Xue Y, Wen C, Marx L, Huang B (2012) An analysis of the nonstationarity in the bias of sea surface temperature forecasts for the NCEP Climate Forecast System (CFS) version 2. Mon Weather Rev 140(9):3003–3016

Ma H-Y, Xie S, Klein S, Williams K, Boyle J, Bony S, Douville H, Fermepin S, Medeiros B, Tyteca S et al (2014) On the correspondence between mean forecast errors and climate errors in CMIP5 models. J Clim 27(4):1781–1798

MacLachlan C, Arribas A, Peterson KA, Maidens A, Fereday D, Scaife AA, Gordon M, Vellinga M, Williams A, Comer RE, et al (2014) Global seasonal forecast system version 5 (GloSea5): a high resolution seasonal forecast system. Q J R Meteorol Soc 141(689):1072–1084

Magnusson L, Alonso-Balmaseda M, Corti S, Molteni F, Stockdale T (2013) Evaluation of forecast strategies for seasonal and decadal forecasts in presence of systematic model errors. Clim Dyn 41(9–10):2393–2409

Murphy JM, Sexton DM, Barnett DN, Jones GS, Webb MJ, Collins M, Stainforth DA (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. Nature 430(7001):768–772

Piani C, Frame DJ, Stainforth DA, Allen MR (2005) Constraints on climate change from a multi-thousand member ensemble of simulations. Geophys Res Lett 32(23). doi:10.1029/2005GL024452

Randall DA, Wood RA, Bony S, Colman R, Fichefet T, Fyfe J, Kattsov V, Pitman A, Shukla J, Srinivasan J, et al (2007) Climate models and their evaluation. In: Climate change 2007: the physical science basis. Contribution of working group I to the fourth assessment report of the IPCC (FAR), pp. 589–662. Cambridge University Press

Rodwell M, Palmer T (2007) Using numerical weather prediction to assess climate models. Q J R Meteorol Soc 133(622):129–146

Sanderson BM, Knutti R, Aina T, Christensen C, Faull N, Frame D, Ingram W, Piani C, Stainforth DA, Stone D et al (2008) Constraints on model response to greenhouse gas forcing and the role of subgrid-scale processes. J Clim 21(11):2384–2400

Severijns CA, Hazeleger W (2005) Optimizing parameters in an atmospheric general circulation model. J Clim 18(17):3527–3535

Smith DM, Eade R, Pohlmann H (2013) A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction. Clim Dyn 41(11–12):3325–3338

Smith DM, Murphy JM (2007) An objective ocean temperature and salinity analysis using covariances from a global climate model. J Geophys Res 112(C2). doi:10.1029/2005JC003172

Smith DM, Cusack S, Colman AW, Folland CK, Harris GR, Murphy JM (2007) Improved surface temperature prediction for the coming decade from a global climate model. Science 317(5839):796–799

Uppala SM, Kållberg P, Simmons A, Andrae U, Bechtold V, Fiorino M, Gibson J, Haseler J, Hernandez A, Kelly G et al (2005) The ERA-40 re-analysis. Q J R Meteorol Soc 131(612):2961–3012

Vannière B, Guilyardi E, Madec G, Doblas-Reyes FJ, Woolnough S (2013) Using seasonal hindcasts to understand the origin of the equatorial cold tongue bias in CGCMs and its impact on ENSO. Clim Dyn 40(3–4):963–981

Visbeck M, Marshall J, Haine T, Spall M (1997) Specification of eddy transfer coefficients in coarse-resolution ocean circulation models*. J Phys Oceanogr 27(3):381–402

Vitart F (2004) Monthly forecasting at ECMWF. Mon Weather Rev 132(12):2761–2779

Wan H, Rasch PJ, Zhang K, Qian Y, Yan H, Zhao C (2014) Short ensembles: an efficient method for discerning climate-relevant sensitivities in atmospheric general circulation models. Geosci. Model Dev 7(5):1961–1977. doi:10.5194/gmd-7-1961-2014

Williams K, Bodas-Salcedo A, Déqué M, Fermepin S, Medeiros B, Watanabe M, Jakob C, Klein S, Senior C, Williamson D (2013) The Transpose-AMIP II experiment and its application to the understanding of Southern Ocean cloud biases in climate models. J Clim 26(10):3258–3274

Williamson D, Blaker AT, Hampton C, Salter J (2015) Identifying and removing structural biases in climate models with history matching. Clim Dyn 45(5–6):1299–1324

Yamazaki K, Rowlands DJ, Aina T, Blaker AT, Bowery A, Massey N, Miller J, Rye C, Tett SF, Williamson D et al (2013) Obtaining diverse behaviors in a climate model without the use of flux adjustments. J Geophys Res 118(7):2781–2793

Zhang S (2011) A study of impacts of coupled model initial shocks and state-parameter optimization on climate predictions using a simple pycnocline prediction model. J Clim 24(23):6210–6226