An Empirical Benchmark for Decadal Forecasts of Global Surface Temperature Anomalies

Matthew Newman

CIRES, University of Colorado, and Physical Sciences Division/NOAA Earth System Research Laboratory, Boulder, Colorado

Email: matt.newman@noaa.gov

Journal of Climate, expedited, submitted 7/31/2012
Revision: 12/31/2012
Accepted: 1/22/2013

ABSTRACT

The suitability of an empirical multivariate red noise (AR1) model, or linear inverse model (LIM), as a benchmark for decadal surface temperature forecast skill is demonstrated. Constructed from the observed simultaneous and one-year lag covariability statistics of annually-averaged sea surface temperature (SST) and surface (2m) land temperature global anomalies during 1901-2009, the LIM has hindcast skill for leads 2-5 and 6-9 years comparable to and sometimes even better than skill of the CMIP5 model hindcasts initialized annually over the period 1960-2000, and has skill far better than damped persistence (e.g., a local univariate AR1 process). Over the entire post-1901 record, the LIM skill pattern is similar but has reduced amplitude. Pronounced similarity in geographical variations of skill between LIM and CMIP5 hindcasts suggests similarity in their sources of skill as well, supporting additional evaluation of LIM predictability. For forecast leads above 1-2 years, LIM skill almost entirely results from three non-orthogonal patterns, one corresponding to the secular trend and two more, each with about ten year decorrelation time scales but no trend, that represent most of the predictable portions of the Atlantic Multidecadal Oscillation (AMO) and Pacific Decadal Oscillation (PDO) indices, respectively. As found in previous studies, the AMO-related pattern also contributes to multidecadal variations in global mean temperature, and the PDO-related pattern has maximum amplitude in the west Pacific and represents the residual after both interannual and decadal ENSO variability are removed from the PDO time series. These results suggest that current coupled model decadal forecasts may not yet have much skill beyond that captured by multivariate red noise.

1. Introduction

The decadal prediction problem has been in an embryonic stage for decades. To progress, we could simply apply the climate community's long experience in understanding seasonal-to-interannual variability and improving its prediction to decadal time scales. For example, a wide array of both physical and empirical methods has been used to make ENSO forecasts (e.g., review by Latif et al. 1998). Statistical forecasts are relatively easy and economical to perform, and their skill can be sufficiently high that they are useful both in their own right and as benchmarks for more complex numerical models (e.g., Livezey 1999; van Oldenborgh et al. 2005; Laepple et al. 2008; Krueger and von Storch 2011).

It seems reasonable then that a similar two-pronged approach of physical and empirical methods could advance decadal prediction. This is not to say that improvement will or can occur as readily as for seasonal forecasting. One concern is that, while ENSO provides a well-defined interannual phenomenon understandable as the result of a defined mechanism (e.g. delayed oscillator theory, recharge-discharge mechanism), there do not yet appear to be so clearly defined decadal phenomena, at least in the Pacific. Large-scale patterns such as the Pacific decadal oscillation (PDO; Mantua et al. 1997) do not dominate decadal variability to the same degree as ENSO dominates interannual variability, and moreover may represent the superposition and/or convolution of a few mechanisms (e.g., Schneider and Cornuelle 2005; Newman 2007), including the low-frequency or reddened tail of interannual phenomena (e.g., Newman et al 2003b; Vimont 2005), rather than the result of one identifiable physical process. The effects of anthropogenic climate change complicate comparison between models and observations,

and how to distinguish natural from anthropogenically-forced decadal variability is a fundamental problem (Solomon et al. 2011).

Currently a number of modeling centers have carried out decadal "hindcasts" as part of CMIP5 (Taylor et al. 2012). It is an important long-range goal of climate diagnosis to provide insights that will help improve decadal forecasts from these CGCMs. Here, we diagnose annual to decadal variability and predictability, both unforced and forced, with an empirically determined linear model of the observed system.

2. Multivariate red noise

Climate variability is often characterized by a notable separation between the dominant time scales of interacting processes. For example, compared to much longer ocean timescales, weather varies so rapidly that it has almost no memory. Weather forcing of the ocean can then be approximated as white noise forcing of a damped integrator, or univariate red noise for an anomaly scalar time series, the simplest null hypothesis of climate (Hasselmann 1976). When extended to the more general case of anomalies representing many evolving regional patterns of climate variables, this approximation based on time scale separation becomes *multivariate red noise*. As opposed to its univariate counterpart, multivariate red noise contains both stationary and propagating anomaly patterns (so that scalar indices derived from it can have spectral peaks) and allows for non-symmetric dynamical relationships (so that despite the lack of exponential modal instability, some anomalies experience significant but transient growth and evolution over finite time intervals). A nonlinear system usefully approximated by multivariate red noise can be said to be "predictably" linear.

Linear inverse models (LIMs; Penland and Sardeshmukh 1995), which empirically determine multivariate red noise from observations, provide excellent approximations of observed Pacific SST anomaly evolution on time scales ranging from weeks to years. Newman (2007) (hereafter N07) found that a LIM constructed from annually-averaged tropical and North Pacific SSTs reproduced observed tropical-North Pacific relationships on decadal time scales better than most CMIP3 coupled GCMs. Subsequent studies have had similar success in the Atlantic (Hawkins and Sutton 2009, Zanna 2012) and both ocean basins (Vimont 2012).

In this paper, the N07 analysis is extended to a state vector constructed from global SSTs and surface land temperatures. The resulting LIM is shown to have skill comparable to three CMIP5 decadal hindcast models that used yearly start dates for the period 1960-2000. The sources of this skill are diagnosed, and evaluated in the context of simpler climate indices.

3. Data and model details

Datasets used in this study were SSTs from the Hadley Sea Ice and Sea Surface Temperature analysis (HadISST; Rayner et al. 2003) and surface (2m) land temperatures from the University of East Anglia Climatic Research Unit (CRU) TS 3.1 dataset (Mitchell and Jones 2005) during the period 1901–2009. Monthly data were interpolated onto 2º latitude x 5º longitude gridboxes. Anomalies were determined by removing the climatological monthly mean from data that was temporally smoothed with a 12-month running mean. This aids our analysis, which does not consider seasonality. However, seasonality is likely still relevant to decadal variability (e.g., Newman et al. 2003b; Vimont 2005). Data were prefiltered in an EOF space retaining about 81% of the SST

and 62% of the surface land temperature variances. SST EOFs were determined within ice-free regions only; the remaining ocean regions were regressed onto the SST PCs to provide complete spatial coverage.

The multivariate AR1 process for an anomaly state vector **x,**

$$\mathbf{x}(t+1) = \mathbf{G}_1\mathbf{x}(t) + \sigma(t), \tag{1}$$

is the integrated solution of the dynamical system

$$\frac{d\mathbf{x}}{dt} = \mathbf{L}\mathbf{x} + \boldsymbol{\xi} \tag{2}$$

forced by white noise $\boldsymbol{\xi}$, where $\mathbf{G}_1 = exp\ (\mathbf{L})$ and t is measured in years. N07 determined (2) from Pacific SSTs and we have likewise determined it from global SSTs (not shown). When including surface land temperatures, however, a few eigenmodes of $\mathbf{L}$ decorrelate too rapidly to be sampled at 1-year intervals, so here we take the simpler route of instead determining (1) from observations. Note, however, that from *both* (1) and (2) the best forecast $\hat{\mathbf{x}}(n)$ from initial conditions $\mathbf{x}(0)$ for a lead of n years is

$$\hat{\mathbf{x}}(n) = [\mathbf{G}_1]^n\mathbf{x}(0), \tag{3}$$

and the lag covariance statistics of $\mathbf{x}$ over a n year interval is

$$\mathbf{C}(n) = [\mathbf{G}_1]^n\mathbf{C}(0) \tag{4}$$

where $\mathbf{C}(n) = \left\langle \mathbf{x}(t+n)\mathbf{x}(t)^T \right\rangle$ and $\mathbf{C}(0) = \left\langle \mathbf{x}(t)\mathbf{x}(t)^T \right\rangle$. This allows us to still make forecasts using the LIM and to test its assumption of predictably linear dynamics, and the eigenmodes of $\mathbf{L}$ are still relevant since they are also the eigenmodes of $\mathbf{G}_1$.

The leading 11/6 EOFs of anomalous SSTs/surface land temperatures were retained for the model, with the corresponding principal components (PCs) defining a 17-component state vector $\mathbf{x}(t)$.

Finally, the LIM must be tested on independent data, so estimates of $\mathbf{G}_1$ and of hindcast skill were cross-validated: the data record was sub-sampled by withholding 10% of it; then anomalies, EOFs and $\mathbf{G}_1$ were recalculated from the remaining 90%; finally, hindcasts were generated for the independent period. The procedure was repeated for the entire record. LIM hindcast skill was determined from these jack-knifed hindcasts, with hindcasts verified against the complete (untruncated in EOF space) gridded observations. EOF truncation was chosen to maximize globally averaged squared anomaly correlation skill of 1-2 yr cross-validated hindcasts, but results are fairly insensitive to this parameter choice. Additionally, linearly detrended data from the entire record was used with the same EOF truncation (albeit recalculated EOFs) to construct a second "detrended LIM".

Hindcasts from the LIM are compared to hindcasts from three CMIP5 CGCMs: HadCM3 (DePreSys), MPI-ESM-LR, and GFDL-CM2p1. These models were chosen since they were the only available models whose hindcasts were initialized yearly rather than every five years. Skill was determined from the bias-corrected ensemble mean for each hindcast initialization.

4. Results

4.1 Testing the empirical model

We first test the assumption of linear dynamics underlying the LIM. Figure 1 shows the observed lag-autocovariance for n = 2, 4, 6, and 8 years compared to that predicted by (4). Generally, the match is quite good and confirms that the LIM reproduces the statistics of evolving surface temperature anomalies over the 20th century, with deficiencies over Northern Europe and some parts of North America. Also, the LIM

captures the 2-year lag anti-correlation in the tropical eastern Pacific but underestimates its amplitude, likely because some subsurface ocean anomalies are needed in addition to SST to completely represent ocean state evolution within the LIM on time scales greater than a year (Newman et al. 2011). Still, the LIM does *implicitly* include those subsurface effects that are linearly related to SST (Penland and Sardeshmukh 1995), in contrast to a physical dynamical model in which the evolution of the state vector is governed only by *explicitly* represented interactions among its components.

4.2 Skill of LIM and CMIP5 decadal hindcasts

Figure 2 shows skill measured by local anomaly correlation for hindcasts averaged over leads of 2-5 (left panels) and 6-9 (right panels) years. The top two rows show that the LIM sets a much higher benchmark for skill than does damped persistence (i.e., a grid-space univariate AR1 model). Additionally, for yearly start dates from 1960-2000, skill of LIM decadal hindcasts is comparable to and sometimes better than skill from CMIP5 decadal hindcasts. Maxima (e.g., tropical Indian and Atlantic oceans, central North Atlantic, southwestern US, east central Asia) and minima (e.g., eastern equatorial and Northeast Pacific, western South America, off US Atlantic coast) of skill often coincide between the LIM and CGCM hindcasts, both those shown here and those documented in other studies (e.g., van Oldenborgh et al. 2012; Kim et al. 2012). LIM can thus serve as a benchmark for decadal forecast skill.

Hindcast skill for the Atlantic Multidecadal Oscillation (AMO) and PDO indices is shown in Fig. 3; for comparison, root-mean-squared error (rmse) is also displayed. Here, the AMO index is the area-weighted North Atlantic mean SST, between 0ºN and 60ºN, minus the global mean SST (Trenberth and Shea 2006; van Oldenborgh et al.

2012), and the PDO index is the projection of SST on the leading EOF of monthly detrended North Pacific SST anomalies between 20ºN and 60ºN (Mantua et al. 1997). As previously found for CMIP5 hindcasts (Guemas et al. 2012; Kim et al. 2012; van Oldenborgh et al. 2012), LIM AMO skill is generally higher than PDO skill, which drops off very rapidly for leads greater than a year. The LIM again clearly provides a more stringent decadal forecast test than persistence (here determined from the indices time series and not the gridded values), except for short lead PDO forecasts, with skill that is generally higher (although not significantly so) than the CGCMs. Even global mean temperature hindcast skill is generally comparable.

4.3 Diagnosing forecast skill

Since the geographical variations of skill are often similar between the LIM and CGCMs hindcasts, it seems possible that the sources of skill for CGCMs are largely the same as for the LIM, despite very great differences in model construction. The LIM's low order and simplicity makes it straightforward to assess surface temperature decadal predictability. For an infinite ensemble of a perfect model, forecast skill measured by the average anomaly correlation $\rho_\infty(n)$ between forecast and verification anomalies is also a function of the forecast signal-to-noise ratio S at lead time n:

$$\rho_\infty(n) = \frac{S(n)^2}{[1+S(n)^2]^{1/2}} \tag{5}$$

(Sardeshmukh et al. 2000). In the LIM, ensemble spread (due to noise) is assumed to converge to climatology with increasing n but to be state-independent; so on average S is directly related to stronger predictable signal determined from (3) (Newman et al. 2003a). For leads under about 3 years, transient anomaly growth due to interference of the non-

orthogonal eigenmodes is important, but at longer leads forecast skill is higher for those initial states that project more strongly on the least-damped eigenmodes of **L**.

The three leading eigenmodes of **L** are shown in Fig. 4 along with their associated projection coefficient time series. The leading eigenmode, stationary with a very long *e*-folding time (eft), appears to represent the global secular trend pattern. The second/third eigenmode pair nominally propagates, but effectively consists of two distinct quasi-stationary patterns since the period is very much greater than the 11-year eft. This eigenmode pair represents decadal variability, primarily over the Atlantic (most energetic phase) and over the Pacific (least energetic phase); it has no overall trend and is also recovered as the leading eigenmode pair for the detrended LIM, albeit with an eft of about 9 yrs. The time series in Fig. 4d is reminiscent of the AMO. Additionally, since Fig. 4c has a large global mean component, its variation gives rise to multi-decadal fluctuations of the global mean temperature that reduce the warming between about 1945 and 1990 and enhance it before and after this period (not shown), consistent with some recent studies (e.g., Ting et al. 2009; DelSole et al. 2011; Wu et al. 2011). In the Pacific, Fig. 4e is very similar to the second leading eigenmode of the N07 LIM, which was the residual of the PDO once ENSO influences were removed. Note that this pattern has relatively larger amplitude in the western North Pacific, a region whose potential importance to PDO-like variability has also been highlighted by earlier studies [e.g., see review by Kwon et al. 2010]. The remaining eigenmodes all have efts shorter than three years, including the “interannual ENSO” (Fig. 5g) and “decadal ENSO” (not shown, but similar to N07) eigenmodes that have efts of 1.5 and 2 yr, respectively. The relatively short decorrelation time scale but high variability of this eigenmode (Fig. 5h) limits

predictability both along the equatorial east Pacific and in the Northeast Pacific (as in N07); in essence, on decadal time scales ENSO is noise (Solomon and Newman 2011).

The impact of these eigenmodes upon hindcast skill over the entire 1901-2009 record was explored by creating several different sets of hindcast initializations, each with nonzero projections on different subsets of the eigenmodes. Figure 5 shows that almost all of the total hindcast skill (top panels of Fig. 5) can be recovered for hindcasts initialized with data projected on the leading three eigenmodes alone (middle panels of Fig. 5). Figure 6 shows the impact of the least damped eigenmodes on climate index skill. As in Fig. 5, almost all skill is retained for hindcasts initialized with the leading three eigenmodes only, except where ENSO eigenmodes also impact PDO skill (Fig. 6a) primarily at shorter leads. Longer-range PDO skill is related to the least energetic phase, while AMO skill (Fig. 6b) is related to the most energetic phase. Global mean temperature skill (Fig. 6c) is mostly due to the leading eigenmode, but also has a contribution from the most energetic phase since it contributes to both the AMO and global mean. Note also that hindcast skill of the "detrended LIM" and full LIM is almost identical for both the AMO and PDO indices; in fact, hindcast skill of both indices is about the same whether or not the leading eigenmode is included in initializations of the full LIM. Apart from these indices, however, much of the LIM skill appears due to the trend (cf. top and bottom panels of Fig. 5), or alternatively the leading eigenmode, as has also been suggested for the CGCMs (van Oldenborgh et al. 2012).

5. Concluding Remarks

A multivariate red noise (AR1) model, empirically constructed from annually averaged surface temperature observations using a one-year lag, has been shown to be a more

suitable benchmark for decadal forecasts than damped persistence. In fact, it appears that CMIP5 CGCM decadal hindcast skill does not notably exceed that expected from a predictably linear dynamical system. To the extent that LIM and CGCM skill are comparable, both in amplitude and in geographical variation, the much simpler LIM can also be used to diagnose sources of forecast skill for both forecast systems.

Estimates of decadal predictability from even 100 years of data are necessarily limited (e.g., Wunsch 2012), so it is important to view the results of this paper with some caution. In the LIM we obtained, virtually all long-range skill comes from the leading three eigenmodes with longest *e*-folding times. The leading eigenmode represents the global secular trend pattern while the next eigenmode pair represents decadal variability. Note that this eigenmode pair does not *propagate* with a multidecadal period, but instead has a sufficiently long *e*-folding time that it varies on a multidecadal timescale. The most notable deficiency in CGCM hindcast skill compared to the LIM appears to be related to this eigenmode over the Pacific. It is interesting that the similar eigenmode found in the Pacific-only LIM was poorly simulated in all the CMIP3 pre-industrial control and historical model simulations (N07; Solomon et al. 2011). Whether the global version of this eigenmode continues to be poorly represented by the CMIP5 models, and if so, why, is a subject for further investigation.

6. Acknowledgements

The author thanks Mike Alexander, Prashant Sardeshmukh, Amy Solomon, and three anonymous reviewers for helpful comments. This work was supported by a grant from NOAA CVP.

References

DelSole, T., M. K. Tippett, and J. Shukla, 2011: A significant component of unforced multidecadal variability in the recent acceleration of global warming. *J. Climate*, **24**, 909–926.

Guemas, V., F. Doblas-Reyes, F. Lienert, Y. Soufflet, and H. Du, 2012: Identifying the causes of the poor decadal climate prediction skill over the North Pacific. *J. Geophys. Res*., doi:10.1029/2012JD018004, in press.

Hasselmann, K., 1976: Stochastic climate models. Part I. Theory. *Tellus*, **28**, 474—485.

Hawkins, E., and R. Sutton, 2009: Decadal predictability of the Atlantic ocean in a coupled GCM: Forecast skill and optimal perturbations using linear inverse modeling. *J. Climate*, **22(**14), 3960–3978.

Kim, H.-M., P. J. Webster, and J. A. Curry, 2012: Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts, *Geophys. Res. Lett.*, **39**, L10701, doi:10.1029/2012GL051644.

Krueger, O., J.-S, Von Storch, 2011: A Simple Empirical Model for Decadal Climate Prediction. *J. Climate*, **24**, 1276–1283. doi: http://dx.doi.org/10.1175/2010JCLI3726.1

Kwon, Y.-O., M.A. Alexander, N.A. Bond, C. Frankignoul, H. Nakamura, B. Qiu, L. Thompson, 2010: Role of Gulf Stream, Kuroshio-Oyashio and their extensions in large-scale atmosphere-ocean interaction: A review. *J. Climate*, **23**, 3249–3281. doi: http://dx.doi.org/10.1175/2010JCLI3343.1

Laepple, T., S. Jewson, and K. Coughlin, 2008: Interannual temperature predictions using the CMIP3 multi-model ensemble mean. *Geophys. Res. Lett*., **35**, L10701, doi:10.1029/2008GL033576.

Latif, M., and and Coauthors, 1998: A review of the predictability and prediction of ENSO. *J. Geophys. Res*., **103**, 14375-14393.

Livezey, R., 1999: The evaluation of forecasts. In Analysis of Climate Variability: Applications of Statistical Techniques, H. von Storch and A. Navarra, Eds., Springer Verlag, 177–196.

Mantua, N. J., S. R. Hare, Y. Zhang, J. M. Wallace, and R. Francis, 1997: A Pacific interdecadal climate oscillation with impacts on salmon production. *Bull. Amer. Meteor. Soc.*, **78**, 1069–1079.

Mitchell, T. D., and P. D. Jones, 2005: An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int. J. Climatol*. **25**, 2005.

Newman, M., 2007: Interannual to decadal predictability of tropical and North Pacific sea surface temperatures. *J. Climate*, **20**, 2333-2356.

Newman, M., P. D. Sardeshmukh, C. R. Winkler, and J. S. Whitaker, 2003a: A study of subseasonal predictability. *Mon. Wea. Rev*., **131**, 1715-1732.

Newman, M., G. P. Compo, and M. A. Alexander, 2003b: ENSO-forced variability of the Pacific Decadal Oscillation. *J. Climate*, **16**, 3853-3857.

Newman, M., M. A. Alexander, and J. D. Scott, 2011: An empirical model of tropical ocean dynamics. *Climate Dynamics*, **37**, 1823-1841.

Oldenborgh, G.J. van, M. Balmaseda, L. Ferranti, T. Stockdale, and D. Anderson, 2005: Evaluation of atmospheric fields from the ECMWF seasonal forecasts over a 15-yr period. *J. Climate*, **18**, 3250–3269.

Oldenborgh, G.J. van, F.J. Doblas-Reyes, B. Wouters and W. Hazeleger, 2012: Skill in the trend and internal variability in a multi-model decadal prediction ensemble. *Clim. Dyn.,* **38**, 1263-1280, doi:10.1007/s00382-012-1313-4.

Penland, C., and P. D. Sardeshmukh, 1995: The optimal growth of tropical sea surface temperature anomalies*. J. Climate*, **8**, 1999—2024.

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

Sardeshmukh, P.D., G. P. Compo, and C. Penland, 2000: Changes in probability associated with El Niño. *J. Climate*, **13**, 4268-4286.

Schneider, N., and B. D. Cornuelle, 2005: The forcing of the Pacific Decadal Oscillation. *J. Climate*, **18**, 4355-4373.

Solomon, A., and the US CLIVAR Working Group on Decadal Predictability, 2011: Distinguishing the roles of natural and anthropogenically forced decadal climate variability: Implications for prediction, *Bull. Am. Meteorol.. Soc*., doi:10.1175/2010BAMS2962.1.

Solomon, A., and M. Newman, 2011: Decadal predictability of tropical Indo-Pacific Ocean temperature trends due to anthropogenic forcing in a coupled climate model. *Geophys. Res. Lett.*, **38**, L02703, doi:10.1029/2010GL045978.

Taylor, K.E., R.J. Stouffer, and G.A. Meehl, 2012: An overview of CMIP5 and the Experiment Design. *Bull. Amer. Meteorol. Soc*., **92**, 485-498, doi: http://dx.doi.org/10.1175/BAMS-D-11-00094.1.

Ting, M., Y. Kushnir, R. Seager, and C. Li, 2009: Forced and Internal 20th Century SST Trends in the North Atlantic. *J. Climate*, **22**, 10.1175/2008JCLI2561.1.

Trenberth, K.E., and D. J. Shea, 2006: Atlantic hurricanes and natural variability in 2005. *Geophys Res Lett*, **33**, L12,704. doi:10.1029/2006GL026894.

Vimont, D. J.. 2005: The contribution of the interannual ENSO cycle to the spatial pattern of decadal ENSO-like variability. *J. Climate*, **18**, 2080-2092.

Vimont, D. J., 2012: Analysis of the Atlantic meridional mode using linear inverse modeling: Seasonality and regional influences. *J. Climate*, **25**, 1194-1212. doi: 10.1175/JCLI-D-11-00012.1

Wu, Z., N. E. Huang, J. M. Wallace, B. V. Smoliak, and X. Chen, 2011: On the time-varying trend in global-mean surface temperature. *Climate Dynamics*, doi:10.1007/s00382-011-1128-8.

Wunsch, C., 2012: Covariances and linear predictability of the Atlantic Ocean. *Deep-Sea Res. II*, in press. http://dx.d oi.org/10.1016/j.dsr2.2012.07.015

Zanna L., 2012: Forecast skill and predictability of observed Atlantic sea surface temperatures. *J. Climate*, **25**, 5047-5056.

7. Figure captions

Fig. 1. Observed (left) and difference between observed and LIM predicted (right) surface temperature lag-covariance for lags of (top) 2 years (middle) 4 years and (bottom) six years. Contour interval is 0.05 K^2; zero contour is removed for clarity and negative values have dashed contours.

Fig. 2. Local anomaly correlation of hindcasts averaged over leads (left) years 2-5 and (right) years 6-9, for the CMIP5 models, damped persistence, and the LIM, for hindcasts initialized yearly from 1960-2000. Contour interval is 0.1 with negative values indicated by blue shading. Shading of positive values starts at 0.1; redder shading denotes larger values of correlation.

Fig. 3. Left panels: anomaly correlation skill comparison for a) PDO and b) AMO indices, and c) global mean temperature, from hindcasts initialized in the years 1960-2000, calculated as described in the text. Using the lag-1 autocorrelation to roughly estimate degrees of freedom, a value of $r \sim (0.4,0.55,0.55)$ is 95% significant (different from 0) for panel (a,b,c). Right panels: same but using root-mean-squared error (rmse) as the skill measure for d) PDO, e) AMO, and f) global mean temperature. Damped persistence is determined from the lag-1 autocorrelation of the observed index time series, not from local gridded values (as in Fig. 2) which would yield lower index skill.

Fig. 4. (a-f) Leading empirical eigenmodes, with their associated projection coefficient time series (determined by the projection of the data on the corresponding adjoint). (g-h) Most energetic phase of eigenmode pair 9/10, which corresponds to “interannual ENSO,” and its time series (least energetic phase not shown). Contour interval is the same in all

left panels. Sign is arbitrary but is consistent with coefficient time series. Red (blue) shading represents positive (negative) values. Also shown are the *e*-folding times (eft) for the eigenmodes, and period (T) for the propagating eigenmodes. Note that in general these eigenmodes do not correspond to the EOFs, which are constrained to be orthogonal.

Fig. 5. Top: LIM hindcast skill for the 1901-2009 period for hindcasts averaged over leads of (left) 2-5 years and (right) 6-9 years. Middle: Difference between the skill of the LIM hindcasts for the 1901-2009 period (i.e., top panels) and a second set of LIM hindcasts where the projection of the initial conditions on only the three leading eigenmodes (Figs. 4a-f) is retained. Bottom: Detrended LIM hindcast skill for the 1901-2009 period for hindcasts averaged over leads of (left) 2-5 years and (right) 6-9 years (verified against detrended data).

Fig. 6. LIM skill of a) PDO and b) AMO indices, and c) global mean temperature, for hindcasts where different initial conditions are used, for the 1901-2009 period. Also shown in all three panels is the detrended LIM skill. See text for description.

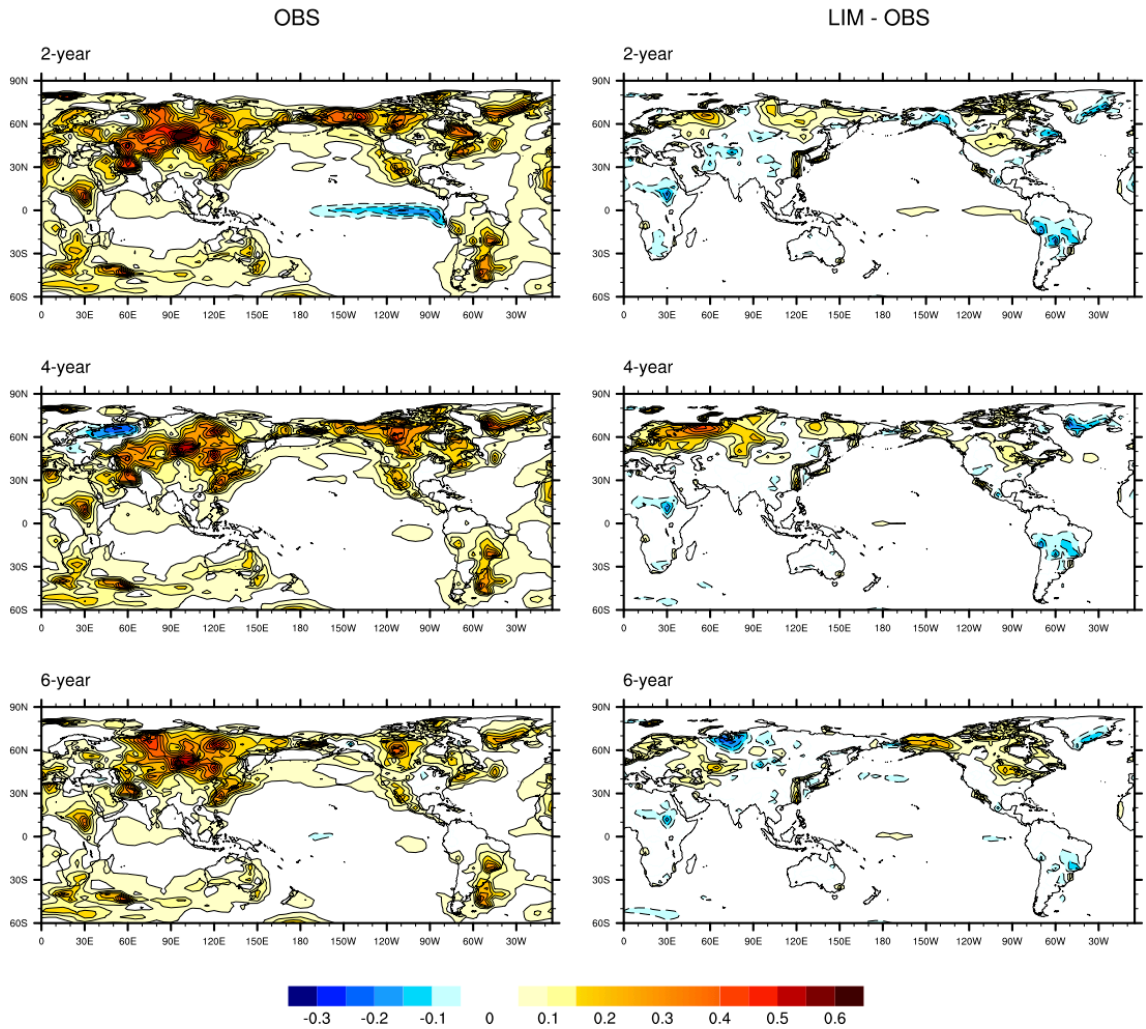

Fig. 1. Observed (left) and difference between observed and LIM predicted (right) surface temperature lag-covariance for lags of (top) 2 years (middle) 4 years and (bottom) six years. Contour interval is 0.05 K^2; zero contour is removed for clarity and negative values have dashed contours.

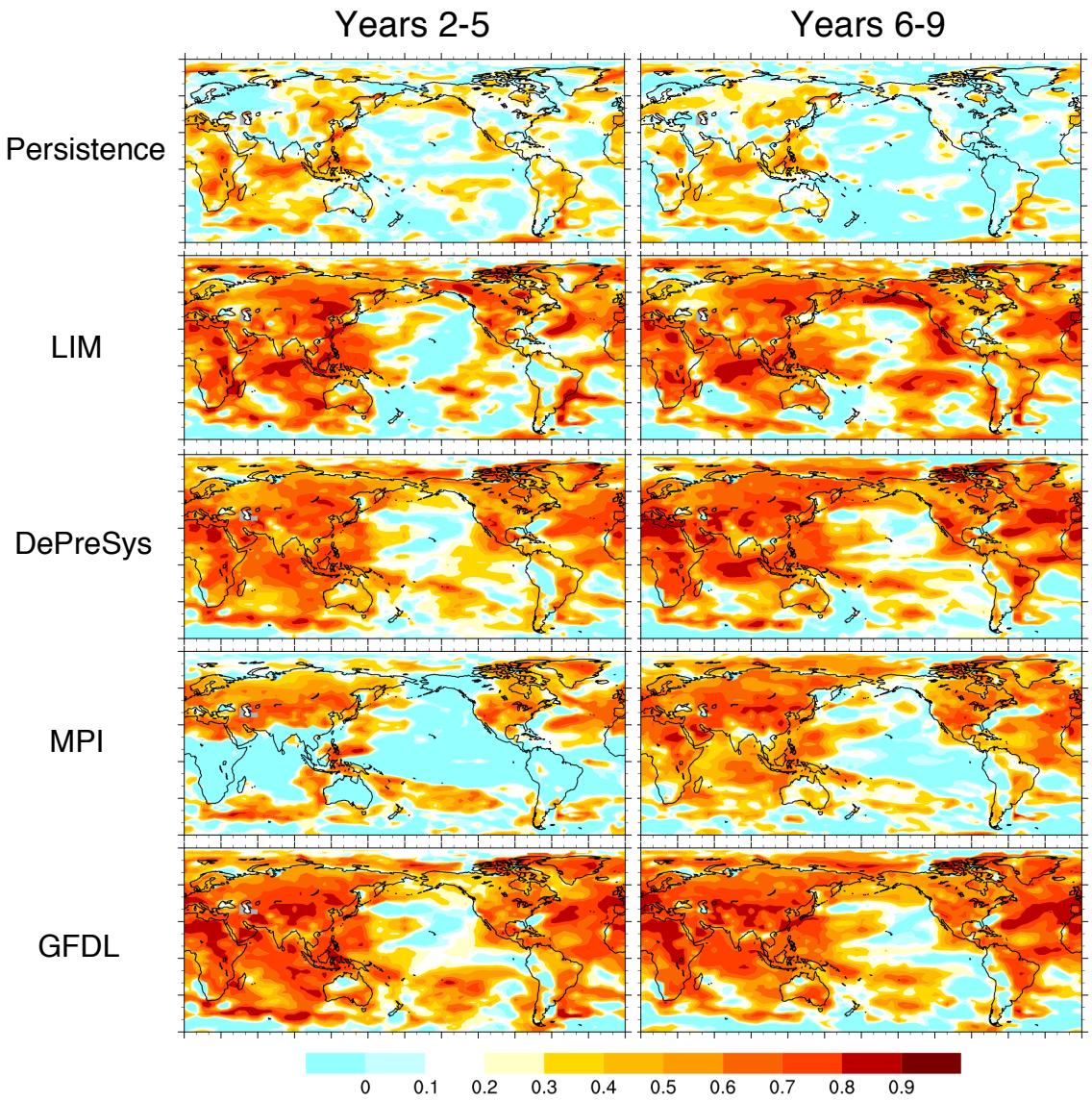

Fig. 2. Local anomaly correlation of hindcasts averaged over leads (left) years 2-5 and (right) years 6-9, for the CMIP5 models, damped persistence, and the LIM, for hindcasts initialized yearly from 1960-2000. Contour interval is 0.1 with negative values indicated by blue shading. Shading of positive values starts at 0.1; redder shading denotes larger values of correlation.

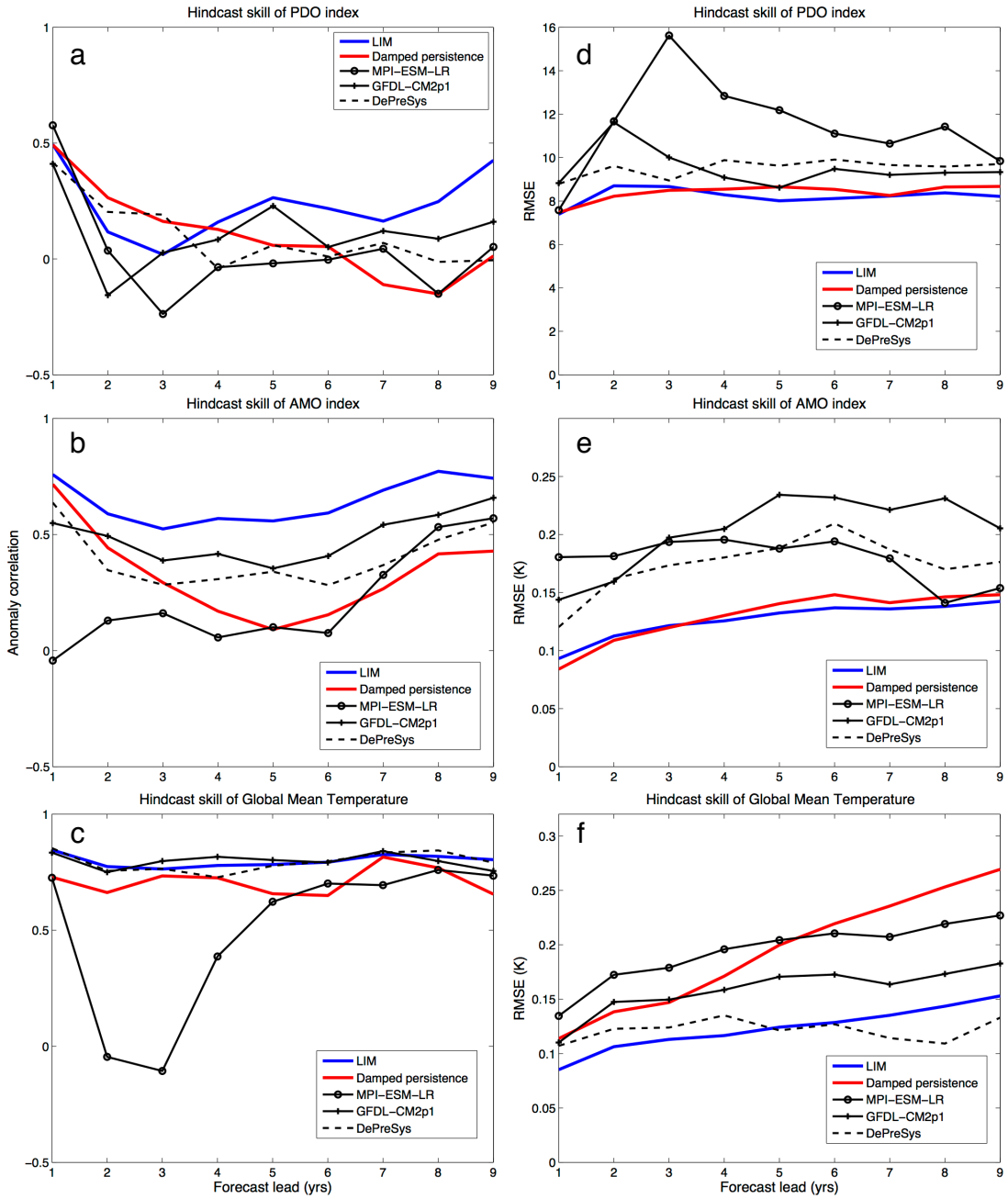

Fig. 3. Left panels: anomaly correlation skill comparison for a) PDO and b) AMO indices, and c) global mean temperature, from hindcasts initialized in the years 1960-2000, calculated as described in the text. Using the lag-1 autocorrelation to roughly estimate degrees of freedom, a value of $r \sim (0.4, 0.55, 0.55)$ is 95% significant (different from 0) for panel (a,b,c). Right panels: same but using root-mean-squared error (rmse) as the skill measure for d) PDO, e) AMO, and f) global mean temperature. Damped persistence is determined from the lag-1 autocorrelation of the observed index time series, not from local gridded values (as in Fig. 2), which would yield lower index skill.

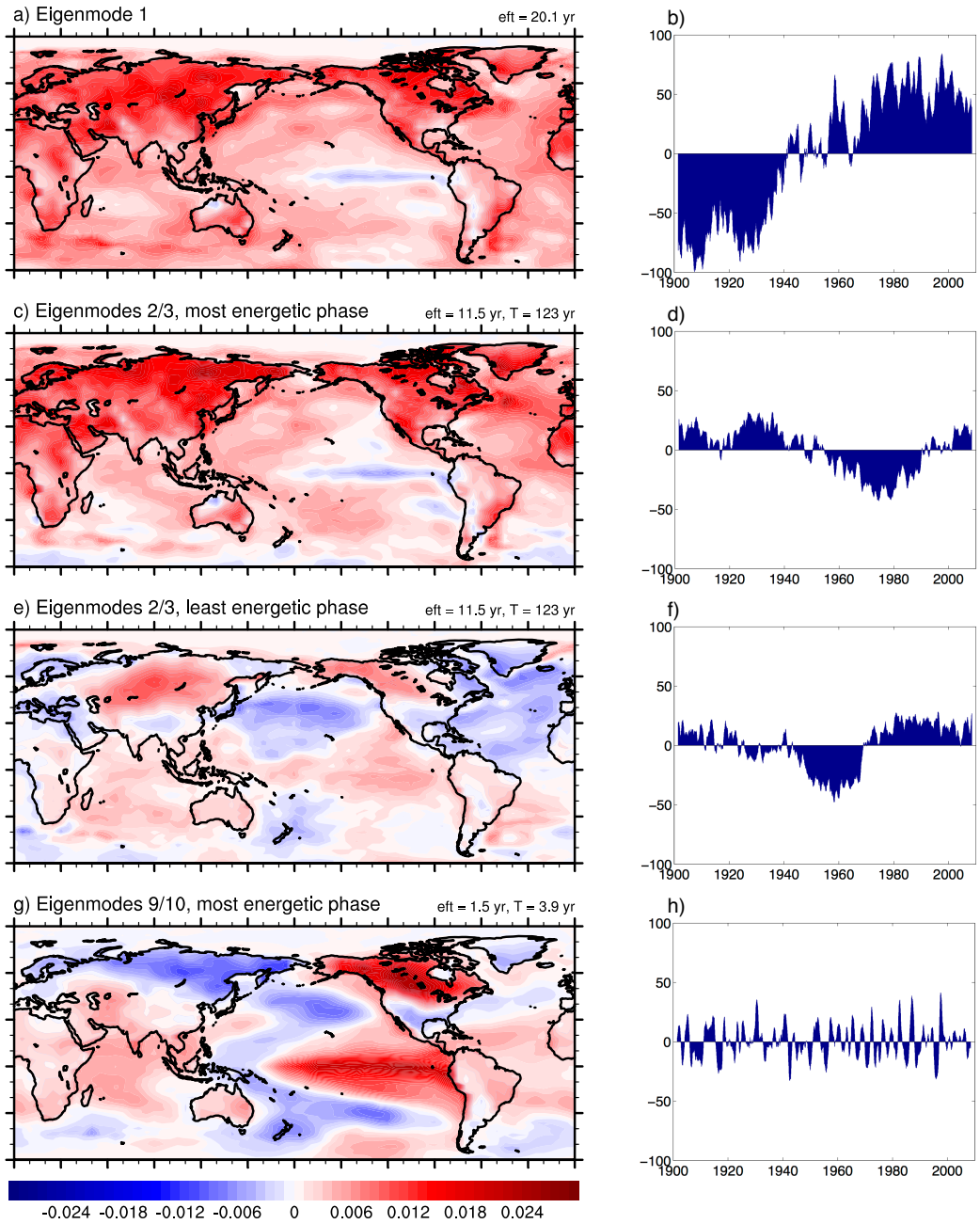

Fig. 4. (a-f) Leading empirical eigenmodes, with their associated projection coefficient time series (determined by the projection of the data on the corresponding adjoint). (g-h) Most energetic phase of eigenmode pair 9/10, which corresponds to "interannual ENSO," and its time series (least energetic phase not shown). Contour interval is the same in all left panels. Sign is arbitrary but is consistent with coefficient time series. Red (blue) shading represents positive (negative) values. Also shown are the *e*-folding times (eft) for the eigenmodes, and period (T) for the propagating eigenmodes. Note that in general these eigenmodes do not correspond to the EOFs, which are constrained to be orthogonal.

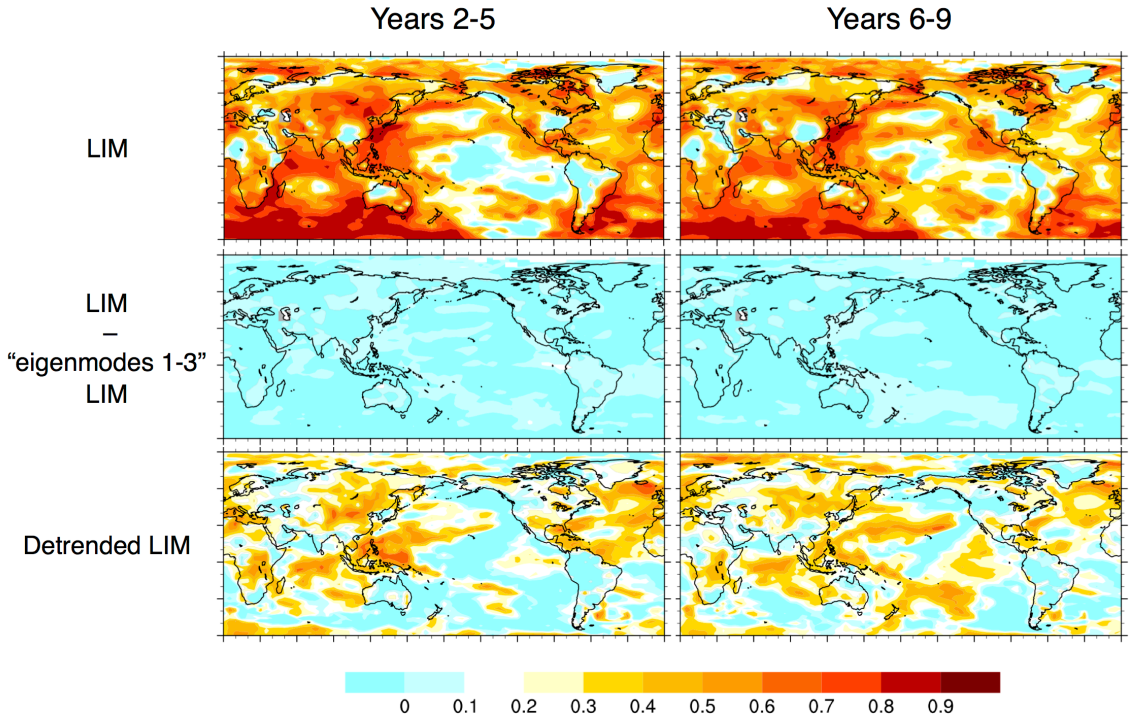

Fig. 5. Top: LIM hindcast skill for the 1901-2009 period for hindcasts averaged over leads of (left) 2-5 years and (right) 6-9 years. Middle: Difference between the skill of the LIM hindcasts for the 1901-2009 period (i.e., top panels) and a second set of LIM hindcasts where the projection of the initial conditions on only the three leading eigenmodes (Figs. 4a-f) is retained. Bottom: Detrended LIM hindcast skill for the 1901-2009 period for hindcasts averaged over leads of (left) 2-5 years and (right) 6-9 years (verified against detrended data).

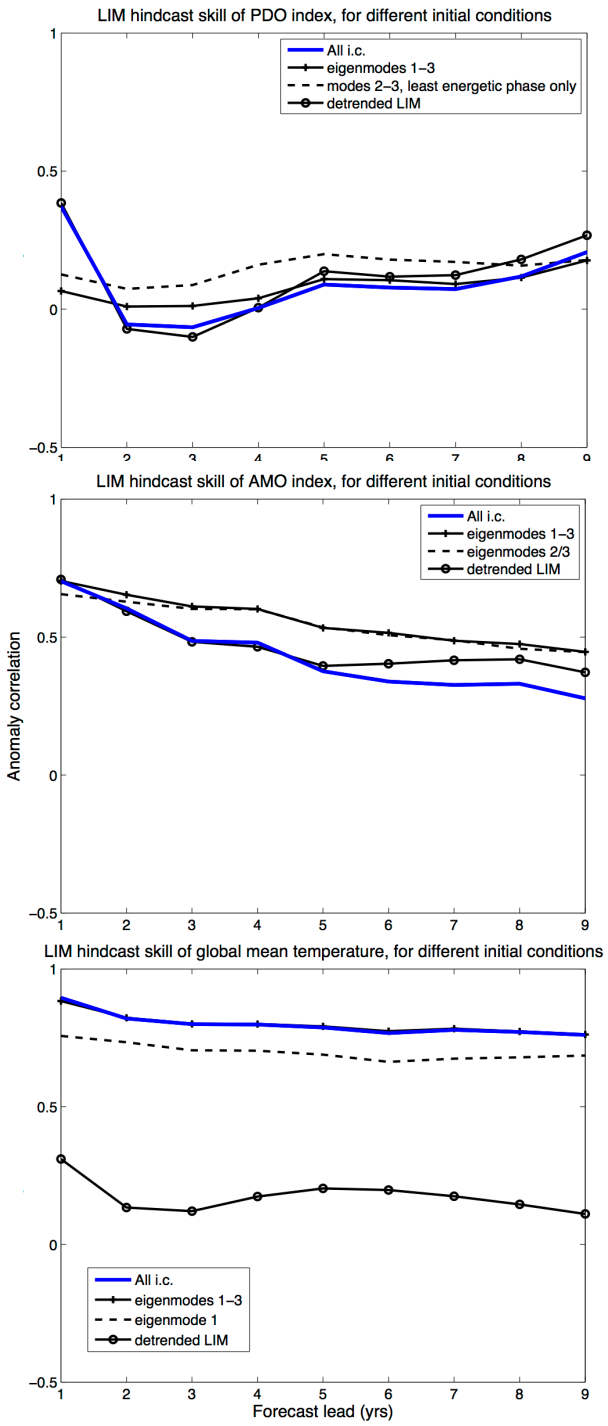

Fig. 6. LIM skill of a) PDO and b) AMO indices, and c) global mean temperature, for hindcasts where different initial conditions are used, for the 1901-2009 period. Also shown in all three panels is the detrended LIM skill. See text for description.