The role of internal variability in seasonal hindcast trend errors

Thomas, R., Woollings, T. and Dunstone, N. (2025) The role of internal variability in seasonal hindcast trend errors. Journal of Climate, 38 (19). pp. 5541-5553. ISSN 1520-0442 doi: 10.1175/jcli-d-24-0367.1 Available at https://centaur.reading.ac.uk/123765/

The Role of Internal Variability in Seasonal Hindcast Trend Errors

RHIDIAN THOMAS,[a,b] TIM WOOLLINGS,[a] AND NICK DUNSTONE[c]

[a] *Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, United Kingdom*
[b] *National Centre for Atmospheric Science, University of Reading, Reading, United Kingdom*
[c] *Met Office Hadley Centre, Exeter, United Kingdom*

(Manuscript received 11 July 2024, in final form 20 June 2025, accepted 13 July 2025)

ABSTRACT: Initialized hindcasts inherit knowledge of the observed climate state, so studies of multidecadal trends in seasonal and decadal hindcast models have focused on the ensemble mean when benchmarking against observed trends. However, this neglects the role of short-time-scale variability in contributing to long-term trends, and hence trend errors. Using a single-model coupled hindcast ensemble, we generate a distribution of 10 000 hindcast trends over 1981–2022 by randomly sampling a single ensemble member in each year. We find that the hindcast model supports a wide range of trends in various features of the large-scale climate, even when sampled at leads of just 1–3 months following initialization. The spread in hindcast global surface temperature trends is equivalent to approximately a sixth of the total observed warming over the same period, driven by large seasonal variability of temperatures over land. The hindcasts also lend support for observed poleward jet shifts, but the magnitude of the shifts varies widely across the ensemble. Our results show that a fair comparison of hindcast trends to observations should consider the full range of model trends, not only the ensemble mean. More broadly, we argue that the hindcast trend distribution offers a largely untapped tool for studying multidecadal climate trends in a very large ensemble, through exploiting existing hindcast data.

SIGNIFICANCE STATEMENT: We show that seasonal forecast models support a wide range of long-term trends in various climate features, from global surface temperature to shifts of the jet streams. This is important because trends in these models are often compared to observed trends to test the model's performance. However, comparisons have typically used the model ensemble mean, neglecting the contribution of short-time-scale variability to long-term trends. We argue that accounting for the full range of model trends is necessary to avoid misdiagnosing trend errors in the models, particularly for features that are sensitive to atmospheric circulation variability, such as regional trends in the extratropics.

KEYWORDS: Climate change; Ensembles; Hindcasts; Internal variability; Trends

1. Introduction

Predictive skill in dynamical models has extended beyond seasonal time scales to include certain features on decadal scales (Kushnir et al. 2019). Although individual simulations typically span only months or a few years, multidecadal trends over sequences of seasonal predictions have also been explored. Differences in long-term trends between hindcast models and observations—referred to as model trend errors—have been linked to low forecast skill in regions where trends differ (Choi et al. 2016; Krakauer 2019; Shao et al. 2021; Becker et al. 2022), including the effect of tropical Pacific trend errors on ENSO prediction (Shin and Huang 2019; L'Heureux et al. 2022; Becker et al. 2022). Recently, trend errors in seasonal hindcasts have also been suggested to shed light on differences between the observed trends and those in freely evolving, uninitialized climate models (Beverley et al. 2024).

An important aspect that has been overlooked is the spread of trends in initialized hindcasts that can arise through internal variability, with each of the cited studies above focusing on the ensemble-mean hindcast trend. In an ensemble of uninitialized climate simulations using a single model, trends in individual realizations will differ from each other due to differences in internally generated variability. Since the phasing of this variability is not coherent across the ensemble, it is averaged out in the ensemble mean, leaving the forced model trend (e.g., Deser et al. 2012).

On the other hand, hindcasts are initialized from observed conditions, so the initial phasing of internal variability in each ensemble member is inherited from the observed state. Slow modes of variability, such as in the ocean circulation, change little over the forecast window and remain coherent across the ensemble. This is shown in Figs. 1c and 1d, which show time series derived from the hindcast dataset that we use in this paper. These are given at both 1-month lead and 7-month lead (blue and green shading, respectively) and compared with the equivalent free-running model (orange). Whereas the phasing of ocean variability varies between members of

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-24-0367.s1.

Corresponding author: Rhidian Thomas, r.h.thomas@reading.ac.uk

DOI: 10.1175/JCLI-D-24-0367.1

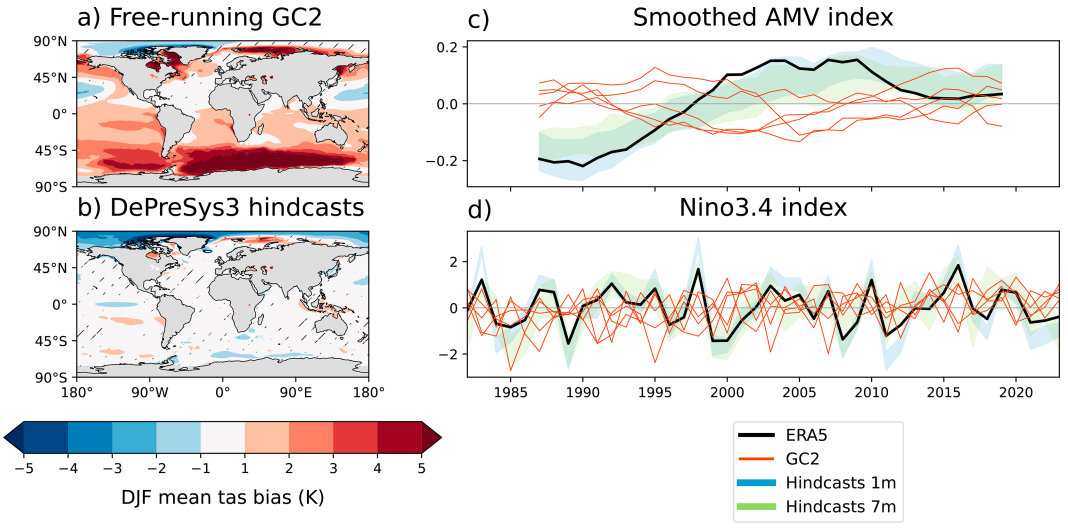

FIG. 1. In the hindcasts, mean state biases are reduced and coupled variability is constrained by the observed system. DJF SAT bias in (a) the free-running model, HadGEM3-GC2, and (b) the DePreSys3 hindcasts at 1-month lead. The bias is calculated as the time-mean difference between the model and ERA5 (model minus reanalysis). The time mean is taken over the same period as the trend calculations (DJF 1981/82–2022/23). Hatching indicates where greater than 20% of the ensemble members disagree on the sign of the bias. Time series plots of (c) the Atlantic multidecadal variability (AMV) index (Trenberth and Shea 2006) and (d) the Nino-3.4 index (Trenberth 1997), each for DJF. ERA5 is in black and the five free-running members are in orange. Green shading indicates the 5%–95% range of values in the hindcast ensemble. A 10-yr centered rolling mean is applied to the AMV index.

the uninitialized ensemble, the hindcasts remain broadly in phase with the observed evolution (black line) at up to 7-month lead time. The hindcast ensemble-mean trend therefore contains both the model's forced trend and a component due to slow coupled modes that is coherent with observed variability.

Deviations of the hindcast ensemble-mean trend from the observed trend have thus been interpreted as model trend errors (Beverley et al. 2024). However, the memory of the initial conditions is lost more quickly in the atmosphere, where small differences at initialization lead to ensemble dispersion at seasonal verification lead times. This short-time-scale variability is not coherent across the hindcast ensemble and is averaged out in the ensemble mean, but it could be an important contributor to observed trends (and hence ensemble-mean trend errors). In this paper, we introduce a method to quantify the full range of trends that is possible in an initialized hindcast ensemble through sampling individual members in each year. We show that variability on seasonal time scales can lead to nontrivial differences in various multidecadal climate trends across the Met Office Decadal Prediction System, version 3 (DePreSys3), hindcast ensemble (Dunstone et al. 2016). This means that a fair comparison to observed trends should use the full distribution of possible hindcast trends, rather than only the ensemble-mean trend.

The main focus of this paper is on the spread of trends in the DePreSys3 prediction system over 1981–2022, a period covering most of the satellite era. We also compare to trends in five freely evolving simulations run using the same model as DePreSys3 (HadGEM3-GC2, see section 2) and covering the same time period. It is relatively uncommon to have hindcasts and directly comparable free-running simulations from the same model, so these provide a valuable comparison even though the ensemble size is limited. The full-field initialization strategy used in DePreSys3 may also reduce the development of some mean-state biases seen in the free-running model (cf. Figs. 1a,b). There is some evidence that mean state biases in the control period can affect the response of a model to external forcing (Simpson et al. 2021), so the reduced biases could lead to improved trends in the hindcasts. These comparisons could be used to identify trends that are especially affected by the initialization in order to target future studies.

Details of the models used and the generation of the trend ensemble are in section 2. Zonal-mean trends in temperature and zonal winds are presented in sections 3 and 4. Section 5 discusses hindcast trends in surface air temperature. Each results section will compare ensemble-mean trends in the hindcasts to trends in observations or reanalyses. Then, by presenting the full ensemble spread, we will demonstrate where comparing only to the ensemble means can give a misleading picture. Finally, concluding remarks are in section 8.

2. Methods

a. *Hindcast data*

Monthly hindcast data are used from DePreSys3, the third iteration of the Met Office decadal prediction system (Dunstone et al. 2016). DePreSys3 uses the HadGEM3-GC2 coupled model (Williams et al. 2015; Senior et al. 2016). HadGEM3-GC2 incorporates the Met Office Unified Model (UM) global atmospheric v6 dynamical core, run here at N216 resolution (0.83° × 0.56°,

approximately 60 km in the midlatitudes). The atmospheric model has 85 levels in the vertical and an 85-km model top with a well-resolved stratosphere. The land surface model is JULES (Best et al. 2011). HadGEM3-GC2 uses the Global Ocean version 5.0 configuration of the NEMO ocean model (Megann et al. 2014), run at an eddy-permitting horizontal resolution of 0.25° and with 75 vertical levels. The sea ice model is CICE (Rae et al. 2015).

Initial conditions are derived from an assimilation run in which the model fields are nudged toward full-field analyses (relaxation time scale in square brackets): in the atmosphere, 6-hourly temperature, zonal winds, and meridional winds, from ERA-Interim (until 2019; Dee et al. 2011)/ERA5 (after 2019; Hersbach et al. 2020) (6 h); monthly salinity and temperature in the ocean from the Met Office Statistical Ocean Reanalysis (MOSORA; Smith et al. 2015) (10 day); and monthly HadISST (Rayner et al. 2003) sea ice concentration (1 day) (Dunstone et al. 2016). During the assimilation run, nonnudged fields (e.g., snow cover and soil moisture) are free to evolve in response to the nudged parameters. This nudging strategy aims to reduce the shock that could arise if initialized using instantaneous fields at the launch time. Simulations are launched biannually on the 1 November and 1 May for every year since 1960, although the work presented here focuses on the data-rich period since the introduction of satellite observations in 1980.

An ensemble of 40 members for each start date is generated by feeding different seeds to a stochastic physics scheme which perturbs the model physics tendencies (Bowler et al. 2009). Simulations last between 13 and 66 months. External forcing in DePreSys3 is time-varying, following CMIP5 historical forcing until 2005 and RCP4.5 thereafter.

b. *Free-running model data*

The hindcasts are compared to five free-running members of the same HadGEM3-GC2 model with historical forcing ("GC2"). These continuous model integrations have identical spatial resolution and external forcing as DePreSys3, differing only in being uninitialized. Members are initialized from different points along the preindustrial control run in 1850. Comparing with these like-for-like simulations isolates the effect of initialization on the modeled trends.

c. *Reanalysis data*

We compare trends in the models to those in three modern atmospheric reanalyses: ERA5 (Hersbach et al. 2020), JRA-55 (Kobayashi et al. 2015), and the Japanese Reanalysis for Three Quarters of a Century (JRA-3Q) (Kosaka et al. 2024). Monthly mean data are used for all datasets between June 1981 and February 2023. ERA5 is retrieved at 1° horizontal grid resolution and the two JRA reanalyses at 1.25°. Although reanalysis trends can be susceptible to biases or discontinuities (see below), we use them here for dynamical consistency between the temperature and circulation trends.

Caution is necessary when interpreting trends in reanalyses due to changes in the assimilated observing systems over time (Bengtsson et al. 2004). In JRA-3Q, different SST forcing datasets are used before and after June 1985: Centennial In Situ Observation-Based Estimates SST, version 2 (COBE-SST2), before and Merged Satellite and In Situ Data Global Daily SST (MGDSST) after (Kosaka et al. 2024). The change in dataset could affect trends that span the changeover date, particularly for variables that are sensitive to the lower boundary conditions (e.g., surface air temperature). We find no evidence of a sharp discontinuity in 1985 for the variables considered here (not shown); additionally, the JRA-55 dataset is included for comparison, which uses the COBE-SST2 forcing dataset throughout (Kobayashi et al. 2015). The overall similarity between trends in JRA-3Q and JRA-55 builds confidence in the use of JRA-3Q.

Kosaka et al. (2024) note that surface air temperature (SAT) observations assimilated in JRA-55 and JRA-3Q over ocean may be influenced by biased ship observations (Simmons et al. 2004); when calculating global mean SAT in the JRA reanalyses, they therefore use analysis fields over land and the background forecast field over ocean. This paper adopts the same approach for the JRA reanalyses. Without this step, the analysis global SAT trends are approximately 20% weaker than in ERA5 (Simmons 2022; not shown).

d. *Trend analysis*

Seasonal-mean trends are presented for DJF and JJA. DJF trends are calculated from DJF 1981/82 to DJF 2022/23, and JJA trends are calculated from 1981 to 2022. Both May and November start dates are exploited to study trends at leads of 1–3 months ("1-m lead") and 7–9 months ("7-m lead").

Multidecadal trends are constructed by randomly selecting a single member for each start date (i.e., one member out of 40 is chosen each year, and these are stitched together to form a time series). Hindcast labeling is arbitrary; no special relationship exists between the ith member for one start date and the ith member for another. This allows a free choice of all 40 members for each start date between 1981 and 2023. The selection process is repeated 10 000 times to generate a distribution of time series consistent with the hindcast predictions (Kelder et al. 2020; Kay et al. 2022). In the following sections, "ensemble" is used to refer to this distribution of 10 000 42-yr sequences ("members"), rather than to the 40-member ensemble for each forecast run.

Linear trends for each of the 10 000 time series are calculated using least squares regression. Significance in the reanalyses is calculated using a two-tailed t test with null hypothesis of zero trend. The test uses an effective sample size that accounts for lag-1 autocorrelation (Santer et al. 2000); as the trends are calculated here using seasonal-mean data, this effect is generally small. To account for multiple testing in the reanalyses, only those regions where the local test statistic passes the false discovery rate criterion (Wilks 2016), with a global control level of $\alpha = 0.1$, are stippled in Figs. 2 and 4. DePreSys3 and GC2 output is conservatively regridded to the coarser 1° resolution for comparison with ERA5. The JRA reanalyses are not regridded as no point-by-point comparison is made with the other datasets. Only data above the surface are used for all pressure-level variables.

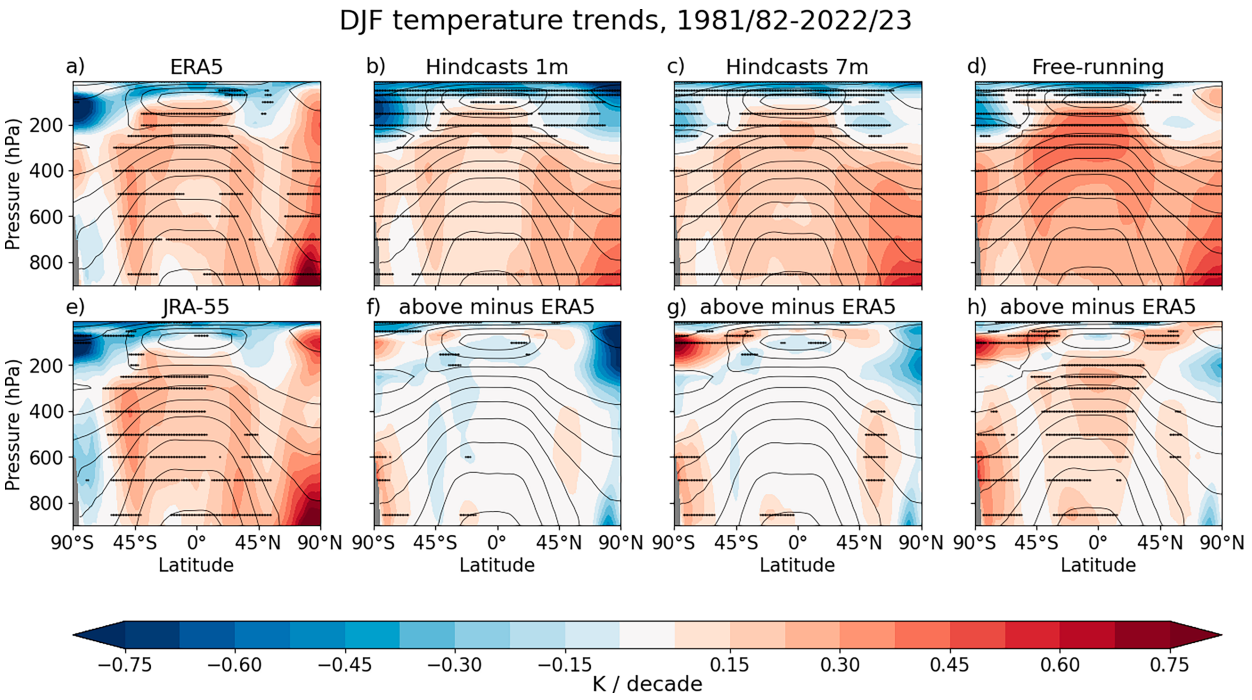

FIG. 2. DJF zonal-mean temperature trends (1981/82–2022/23) for (a) ERA5, the hindcast ensemble mean for (b) 1-m lead and (c) 7-m lead, (d) the free-running GC2 ensemble mean, and (e) JRA-55. Stippling on (a) and (e) indicates a trend significantly different from zero at the 5% level after accounting for multiple testing (see methods). Stippling on (b)–(d) indicates where at least 80% of the model members agree on the sign of the trend. (f)–(h) Differences of the plots in (b)–(d) from the ERA5 trend in (a). Stippling on (f) and (g) indicates where the hindcast mean is more than 2S away from the ERA5 mean, where S is the standard error on the ERA5 regression. On both rows, solid contours show the corresponding climatological values with spacing of 10 K.

3. Zonal-mean temperature

We begin by presenting the zonal-mean temperature trends in ERA5, JRA-55, and the model ensemble means (DJF is shown in Fig. 2 and JJA is shown in Fig. S1 in the online supplemental material). This will provide the context for later sections where the zonal circulation and surface temperature trends are examined in more detail.

Both reanalyses show moderate warming of the troposphere between ±45° (Figs. 2a,e and Figs. S1a,e) and strong warming near the Arctic surface in DJF (Figs. 2a,e). This general structure is captured well by the hindcast ensemble mean at both lead times (Figs. 2b,c and Figs. S1b,c), although the magnitude of the Arctic warming is weaker in the hindcast ensemble mean than the reanalyses. Figures 2f–2h and Figs. S1f–S1h show the ensemble-mean model trend errors relative to ERA5. A striking feature of GC2 is its warm trend bias in the tropical troposphere (Fig. 2h and Fig. S1h), in common with other coupled models (e.g., Mitchell et al. 2020). By contrast, ensemble-mean trend errors in the tropics and midlatitudes are substantially reduced in the hindcasts at both 1-m lead and 7-m lead. The drift over the hindcast runs appears to be small, so that the trend error at 7-m lead is similar to 1-m lead.

The spread of trends in the 1-m hindcasts is shown in Fig. 3, defined as the standard deviation of trends across the ensemble. The spread is relatively small in the tropics, highlighting the tight constraint on the tropical troposphere provided by the initial conditions. By contrast, the spread is much larger in the polar regions, indicating that variability on seasonal time scales has a larger impact on trends at higher latitudes. Importantly, high-latitude regions that show notable trend errors in the hindcast ensemble mean, such as the Arctic lower troposphere and the polar stratosphere in both hemispheres (Figs. 2f,g), also show a wide spread of trends between members (Fig. 3). When we compare reanalysis trends in the Arctic lower troposphere and stratosphere to the hindcasts, the reanalyses are within the 95% confidence intervals of the ensemble (not shown). While the ensemble mean may be sufficient for diagnosing temperature trend errors at low latitudes, Fig. 3 shows that a fair comparison in the extratropics should use the full range of hindcast trends.

4. Zonal-mean zonal winds

Figure 4 shows zonal wind trends in ERA5 and the model ensemble means. In ERA5, the trend over the satellite era is toward a poleward shift of the jets in both hemispheres and seasons (Figs. 4a,e), with the deep structure of the midlatitude wind trends indicating the influence of transient eddies. The exception is the boreal summer jet, where wind trends in the midlatitudes are generally weak (Fig. 4e).

The ensemble-mean hindcast trends are shown in Figs. 4b, 4c, 4f, and 4g. Various features of the ERA5 trend are well

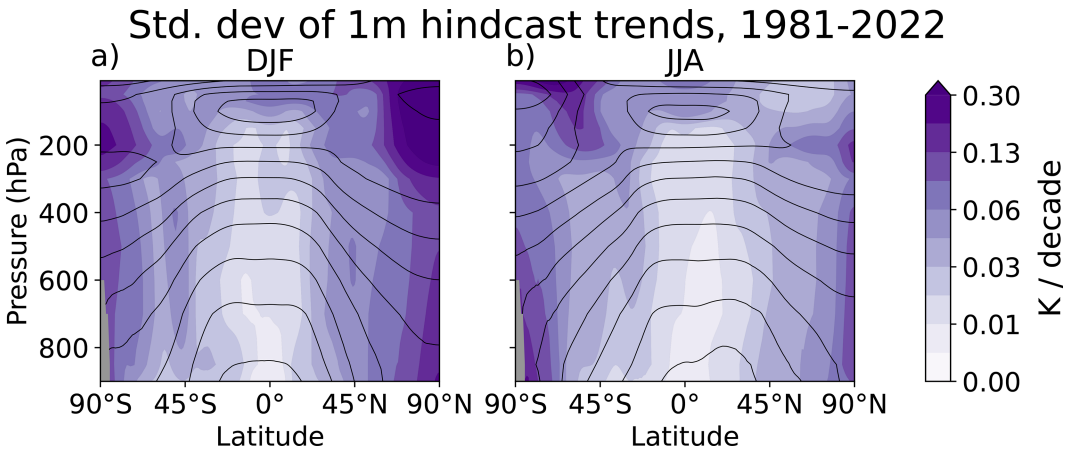

FIG. 3. Spread of zonal-mean temperature trends in the 1-m hindcasts, defined as the standard deviation of trends across the hindcast ensemble. The spread is shown for (a) DJF (1981/82–2022/23) and (b) JJA (1981–2022). Solid contours show the climatological temperature, with spacing 10 K as in Fig. 2. Note the logarithmic color scaling.

reproduced by the 1-m hindcast ensemble mean, such as the strong easterly trends in the subtropics. The deep acceleration on the poleward side of the austral winter jet is also well represented in the hindcast ensemble mean at 1-m lead (Fig. 4f). By comparison, neither the widespread subtropical easterly trends nor the deep westerly trends in austral winter are seen in the free-running ensemble mean (Figs. 4d,h). Even in the ensemble mean, the hindcasts are therefore better able to reproduce reanalysis zonal wind trends.

Next, we discuss the spread of wind trends across the model ensembles. Figure 5 summarizes meridional jet shifts using the zonal indices of Woollings et al. (2023), with positive values indicating poleward shifts. At 1-m lead, the hindcast ensemble means show poleward jet shifts in both hemispheres and seasons,

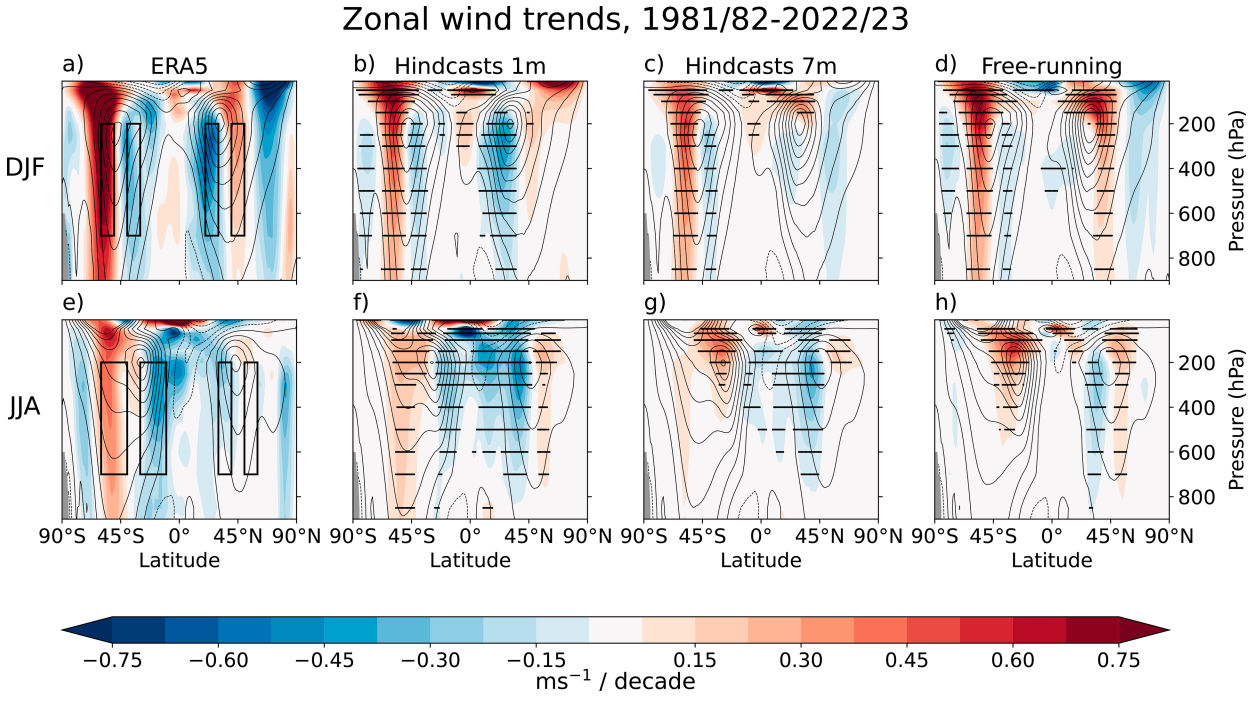

FIG. 4. Zonal-mean zonal wind trends for (a),(e) ERA5, the DePreSys3 hindcast ensemble mean at (b),(f) 1-m lead and (c),(g) 7-m lead, and (d),(h) the free-running ensemble mean (GC2). (top) DJF trends. (bottom) JJA trends. Solid contours show the corresponding climatologies with dashed lines indicating negative values and a spacing of 5 m/s. Stippling follows (a)–(d) in Fig. 2; note that none of the ERA5 grid points show significant trends after controlling for the false discovery rate, as described in the methods. Boxes in (a) and (e) indicate the regions used to calculate the zonal wind indices of Fig. 5. DJF trends are calculated over 1981/82–2022/23, and JJA trends are calculated over 1981–2022.

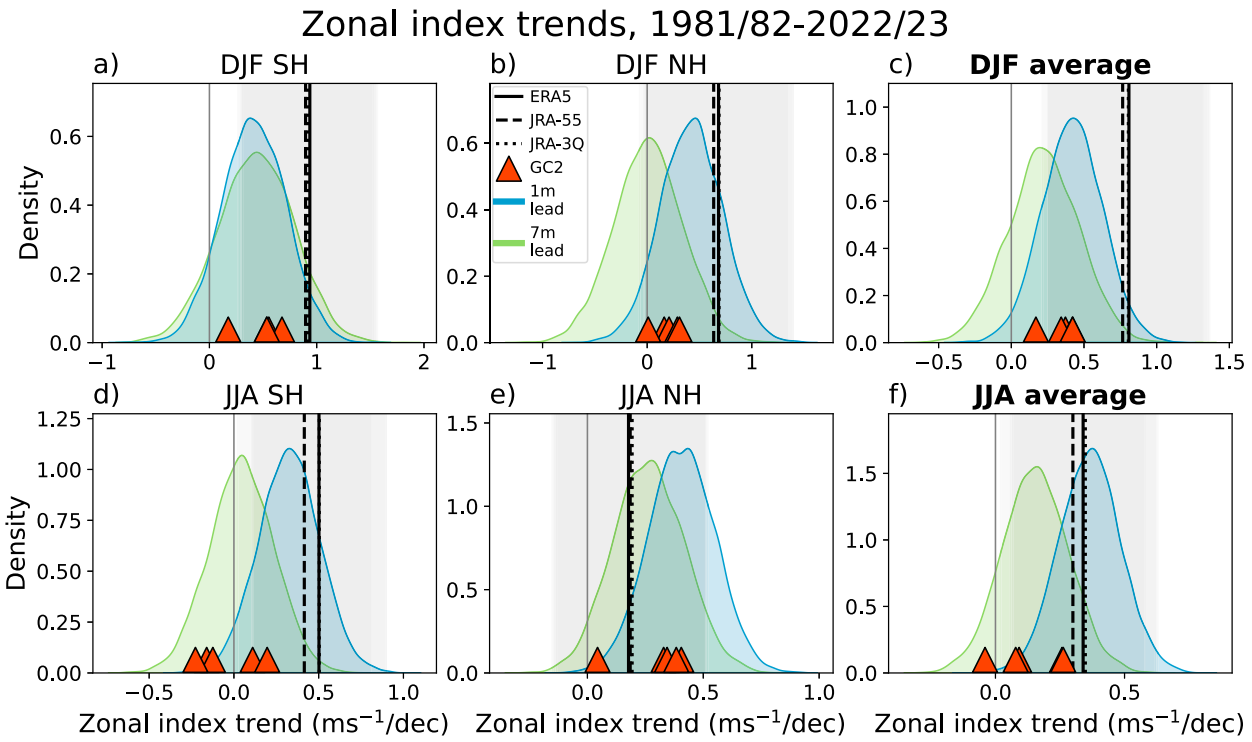

FIG. 5. Jet shifts as measured by trends in the zonal wind indices of Woollings et al. (2023). Each zonal index is the difference in zonal wind trend between the poleward and equatorward sides of the climatological jet, averaged over seven levels between 200 and 700 hPa. The latitude bands used to define each index vary by season and hemisphere, as shown by the boxes in Figs. 4a and 4e. Poleward shifts correspond to positive trends in the index in each hemisphere. Zonal index trends in the (a) SH and (b) NH and (c) the average over both hemispheres for DJF; (d)–(f) as in (a)–(c), but for JJA. Blue and green shading shows the trends in the hindcasts at 1-m lead and 7-m lead, respectively. Reanalysis trends are shown in black, with the gray shading indicating the 95% confidence interval around each product. The free-running GC2 members are shown by the orange triangles.

as seen in Fig. 4. Beyond the ensemble mean, however, a much wider range of trends is possible in the hindcasts. While poleward shifts still dominate, a nonnegligible sample of members in each hemisphere and season exhibits equatorward shifts.

A notable case is austral summer (Fig. 5a). Circulation trends in this season are influenced by changes in stratospheric ozone over the satellite era (e.g., Gillett and Thompson 2003; Son et al. 2010; Seviour et al. 2017). Changes in ozone concentration are nonlinear in time (Banerjee et al. 2020), although the linear trend here is consistent with depletion dominating in the satellite era overall. The ensemble-mean trend in GC2 is a poleward shift of the austral summer jet, consistent with the forced response to ozone depletion. The hindcast jet shift distributions are similar at 1-m lead and 7-m lead, with both showing a poleward shift in the ensemble mean. Even so, 8% of the hindcast trends at 1-m lead (and 11% at 7-m lead) show an equatorward shift of the austral summer jet, despite identical external forcing. In other words, around a tenth of hindcast members show trends in the austral summer jet opposite to the externally forced response due to variability on seasonal time scales. This finding is consistent with the large role for internal variability in austral summer jet shifts identified by Seviour et al. (2017). The hindcasts are somewhat more confident in poleward shifts in the winter jets, though 8% and 4% of the 1-m hindcasts show equatorward shifts in boreal and austral winter, respectively. Even where the hindcasts are relatively confident in the sign of the trend, its magnitude can vary substantially between members.

This has implications when comparing trends in hindcast models with observed trends. The black vertical lines in Fig. 5 denote reanalysis trends, and gray shading indicates the 95% confidence interval on these. In each season and hemisphere, the magnitude of the reanalysis trends differs from the hindcast ensemble means. However, the reanalysis trends clearly fall within the spread of hindcast trends, particularly at 1-m lead, and the reanalysis trends themselves are considerably uncertain as indicated by the wide confidence intervals. Failing to account for either the spread in hindcast trends or the reanalysis trend uncertainty thus gives an incomplete comparison. For example, the mean boreal winter jet shift is weaker in the 1-m hindcasts than in the reanalyses (0.42 m s^{-1} decade^{-1}, compared to 0.68 m s^{-1} decade^{-1} in ERA5, Fig. 5b)—but roughly a quarter of the hindcasts actually shows a stronger poleward shift than in the reanalyses. Diagnosing an equatorward trend error in the model in this case neglects the role that variability on seasonal time scales, largely intrinsic atmospheric variability, can play in influencing multidecadal wind trends.

Despite the wide array of jet shift trends in the hindcasts, some firm conclusions can still be drawn. Averaged over both

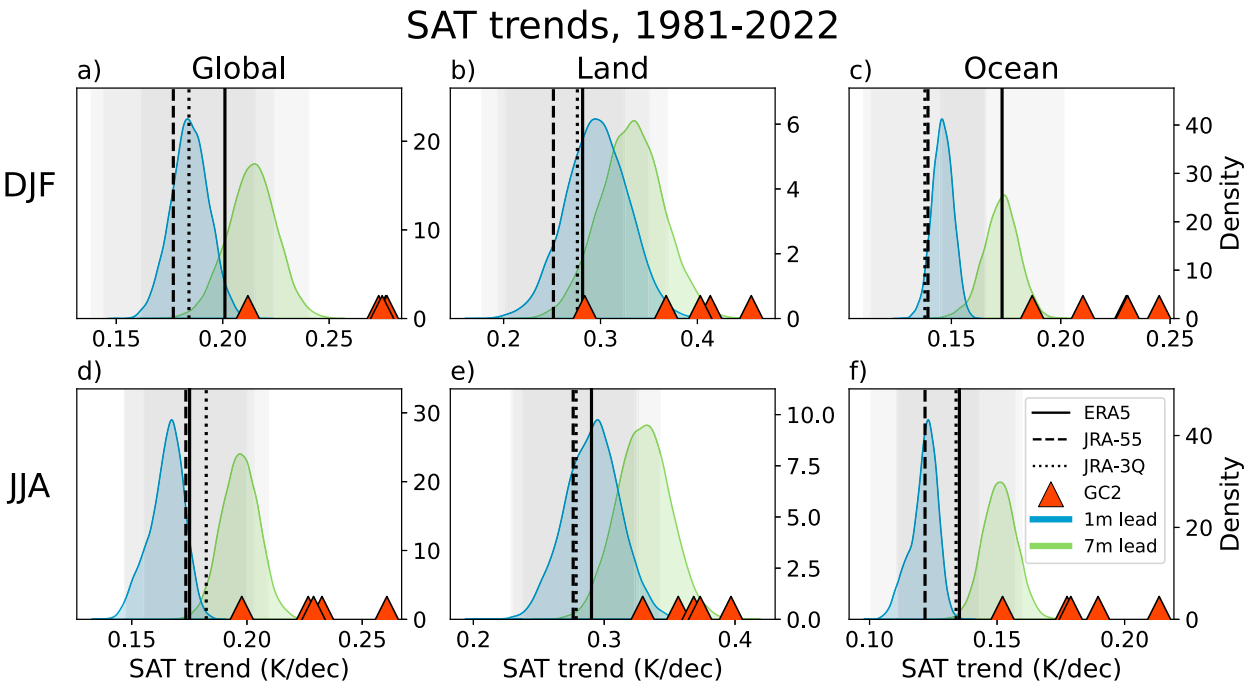

FIG. 6. Area-weighted SAT trends for (top) DJF 1981/82–2022/23 and (bottom) JJA 1981–2022. Trends are averaged (a),(d) globally, (b),(e) over land, and (c),(f) over ocean. Colors, symbols, and shading are as in Fig. 5. Note the different x axes in each panel.

hemispheres, the hindcasts at 1-m lead robustly show a poleward shift of the jets in both DJF and JJA (Figs. 5c,f); only 2% of members fail to show an average poleward shift in DJF and less than 1% in JJA. An average poleward shift globally may hide a weak or equatorward shift in one hemisphere. Accounting for this, we find that 85% of 1-m lead members show concurrent poleward shifts in both hemispheres for DJF and 96% for JJA.

In summary, the DePreSys3 hindcasts broadly lend support to the emergence of poleward zonal-mean jet shifts on the global scale identified by Woollings et al. (2023). However, they also support a wide range of jet shifts for each specific hemisphere and season, including some which are of the opposite sign. The spread in multidecadal hindcast trends arises from variability on seasonal time scales, and failing to account for this can lead to incorrect conclusions regarding trend errors in initialized hindcasts.

5. Surface air temperature

Figures 6a and 6d show trends in averaged global SAT (GSAT) for DJF and JJA, respectively. The reanalysis GSAT trends between 0.17 and 0.20 K decade^{-1} in DJF are well captured by the 1-m lead hindcasts, whereas the reanalysis trends are on the upper edge of the 1-m hindcast distribution in JJA. In both seasons, regression uncertainty on the reanalysis trends is comparable to the spread in hindcast trends over both lead times (not shown). By 7-m lead, the hindcasts drift toward larger trends in each season. Even at 1-m lead, a nonnegligible spread of hindcast GSAT trends is seen, with a 5%–95% range of 0.17–0.20 K decade^{-1} in DJF and 0.15–0.18 K decade^{-1} in JJA. Over the satellite era, this equates to a difference of 0.12 K of warming in DJF and 0.11 K in JJA between hindcast members at the 5th and 95th percentiles, equivalent to around a sixth of the mean warming over the same period or approximately a tenth of the warming since preindustrial times (Gulev et al. 2021).

Figures 6b, 6e, 6c, and 6f show SAT trends averaged over land and ocean, respectively. Initialization of the ocean means that the spread of oceanic SAT trends is relatively narrow in the hindcasts, at both 1-m lead and 7-m lead. The spread is much larger over land in both seasons. The variance in global SAT trends T_{global} is denoted $\sigma^2_{T_{\text{global}}} = \text{Var}(T_{\text{global}})$. This can be expressed as a weighted sum of the trend variances over land $\sigma^2_{T_{\text{land}}}$ and ocean $\sigma^2_{T_{\text{ocean}}}$:

$$\sigma^2_{T_{\text{global}}} = f_L^2\sigma^2_{T_{\text{land}}} + f_O^2\sigma^2_{T_{\text{ocean}}} + 2f_Lf_O\text{Cov}(T_{\text{land}}, T_{\text{ocean}}), \quad (1)$$

where f_L and f_O are the fractional areas of land and ocean, respectively ($f_L + f_O = 1$), and the final term on the right-hand side is the covariance between SAT trends over land and ocean across the ensemble. The relative contributions of each term in the 1-m hindcasts are shown in Table 1. The largest

TABLE 1. Fractions of the total variance in GSAT trends explained by each term in 1 for the 1-m hindcasts. Variances are expressed as fractions of the total variance $\sigma^2_{T_{\text{global}}}$ (i.e., $\sigma^2_{T_{\text{global}}} = 1$).

Season	$f_L^2\sigma^2_{T_{\text{land}}}$	$f_O^2\sigma^2_{T_{\text{ocean}}}$	$2f_Lf_O\text{Cov}(T_{\text{land}}, T_{\text{ocean}})$
DJF	0.87	0.16	−0.03
JJA	0.50	0.27	0.23

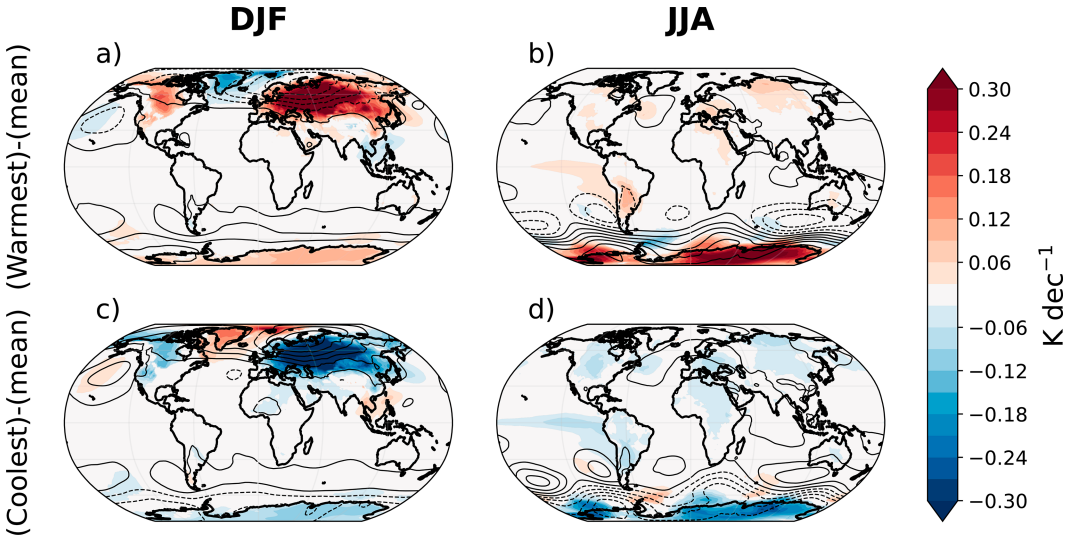

FIG. 7. Difference in SAT and SLP trends for the 200 1-m hindcast members with the (top) warmest and (bottom) coolest GSAT trends in each season. Differences are shown relative to the 1-m hindcast ensemble mean. Only significantly different trends are shaded, using a two-tailed t test at the 5% level. Differences in SLP trend are shown in contours, with solid contours indicating a positive difference in the 200-member subset relative to the ensemble mean.

term in both seasons is the land term, accounting for half the total variance in JJA and the majority in DJF. The ocean contribution is larger in JJA, but it is still only slightly more than half as large as the land term. Finally, while the covariance in land and ocean SAT trends is negligible in DJF, it is reasonably large in JJA, so that SAT trends over land and ocean are positively correlated.

The spatial patterns associated with the spread of GSAT trends are illustrated in Fig. 7. Figure 7a shows the difference in SAT and SLP trends in the 200 1-m hindcast members with the largest DJF GSAT trends minus the ensemble mean. Compared to the ensemble mean, members with the largest GSAT trends are characterized by strong warming of the mid–high latitude continents, particularly Eurasia and North America in the NH. These regional warming differences are consistent with the westerly flow advecting relatively mild oceanic air over the continental landmasses. The spread in hindcast warming over land is large for the same reason that monthly variability in hemispheric temperatures is dominated by land: the heat capacity of the land surface is much smaller than the ocean surface, and hence, its temperature (and the overlying SAT) can adjust much more rapidly in response to changes in the large-scale circulation (Wallace et al. 1996). Wallace et al. (1995) demonstrated that this "cold ocean, warm land" phenomenon explains half of the month-to-month variance in NH SAT anomalies over the 20th century; the spread in multidecadal hindcast trends arises through sampling of this variance.

Figure 7b shows the corresponding plot for JJA. This appears to indicate a greater role for the tropical Pacific than in DJF, consistent with the larger oceanic contribution to the spread in JJA GSAT trends in Table 1. The largest GSAT trends in JJA are particularly associated with warming of high southern latitudes, consistent with greater variability in surface temperatures in the winter hemisphere. In both seasons, small GSAT trends are associated with similar circulation patterns, but opposite in sign (Figs. 7c,d). Hence, while not the dominant factor in the long-term anthropogenically driven GSAT trend, these results highlight that variability on seasonal time scales can noticeably modulate hemispheric and global SAT trends over multidecadal time periods (Iles and Hegerl 2017).

As shown in the previous section for zonal winds, the spread in hindcast trends has implications for diagnosing trend errors using the ensemble mean. Figure 8 shows SAT trends in ERA5 and JRA-55 (Figs. 8a and 8b) and the hindcast ensemble means at 1-m lead and 7-m lead (Figs. 8c,e, respectively). The ensemble-mean hindcast trend errors relative to ERA5 are shown in Figs. 8d and 8f. Figure S2 shows the corresponding plots for JJA. Over ocean, the trend errors are generally small; the spatial correlation between the ERA5 trend and the 1-m lead hindcasts over ocean is 0.89 for DJF and 0.63 for JJA ($p < 0.01$). Many features of the reanalysis trends over ocean are well captured by the hindcast ensemble means, including weak or negative trends in the tropical east Pacific and Southern Ocean (Fig. 8c and Fig. S2c).

On the other hand, the pattern of hindcast trend errors over land generally resembles the GC2 trend errors (cf. Figs. 8d,f and Figs. S2d,f with Fig. 8h and Fig. S2h). The ensemble-mean trend patterns over land remain much smoother in the DePreSys3 hindcasts than in the reanalyses. However, as shown above, this smooth ensemble mean obscures significant regional spread between members in land SAT trends. Several of the land regions with large trend error, such as central Eurasia and northern North America, are also regions with a large spread between members, as shown in Fig. 7. Internal variability is therefore an important factor to consider for SAT trends over land, underlining the importance of using the full range of hindcast trends when diagnosing trend errors relative to observations or reanalyses.

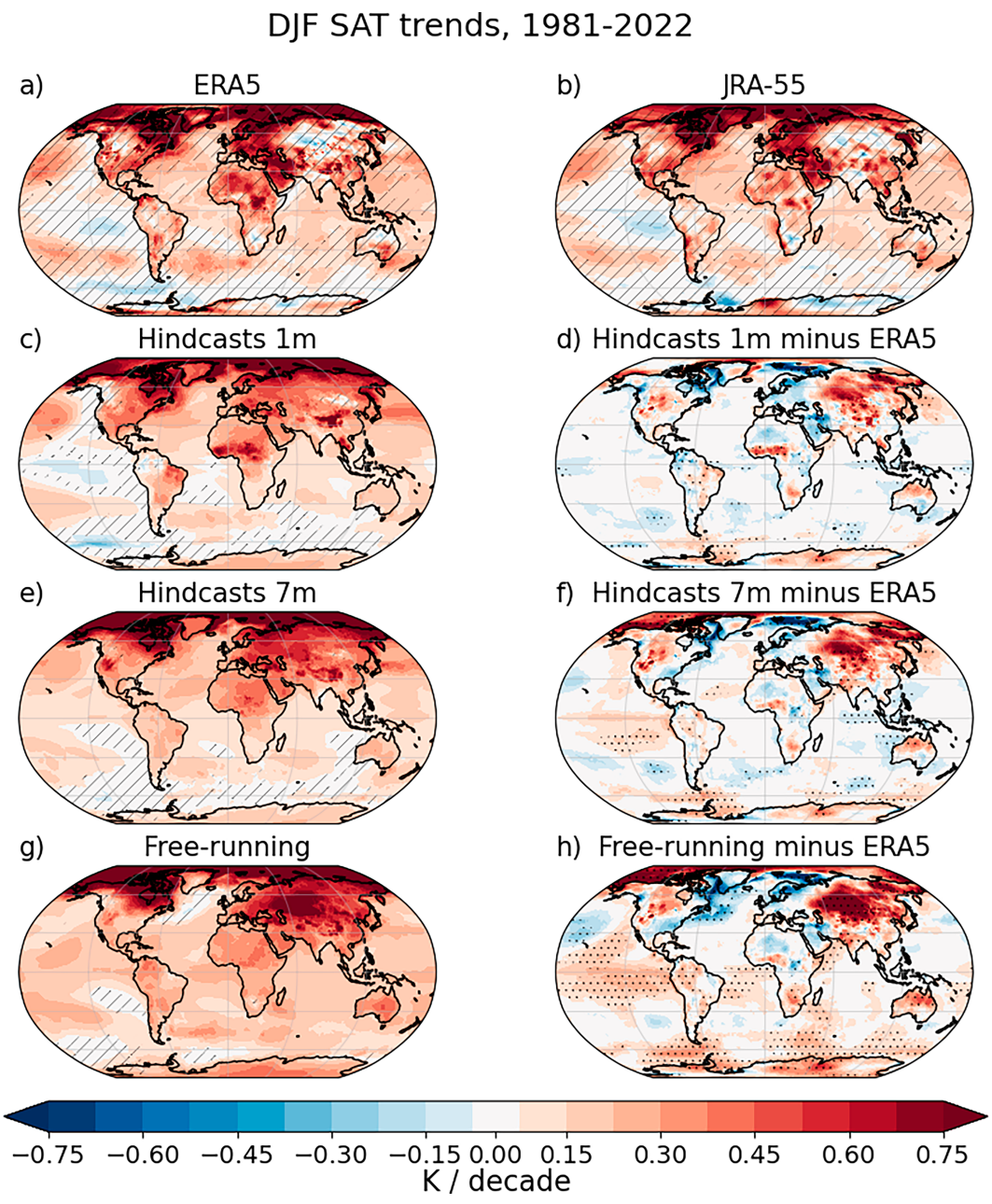

FIG. 8. DJF SAT trends (1981/82–2022/23) for (a) ERA5 and (b) JRA-55, the hindcast ensemble mean at (c) 1-m lead and (e) 7-m lead, and (g) the free-running ensemble mean. Hatching on these plots is the inverse of the stippling on the ensemble-mean trends of Figs. 2 and 4 (here, hatching hides nonsignificant points, defined as before). (d),(f),(h) Differences of the plots in (c), (e), and (g) from the ERA5 trend in (a). As in Fig. 2, stippling on (d), (f), and (h) indicates where the hindcast mean is more than 2S away from the ERA5 mean, where S is the standard error on the ERA5 regression.

6. North Atlantic Oscillation

We now include a brief analysis of trends in the North Atlantic Oscillation (NAO) as an example of a regional circulation structure that has received much attention, both in terms of recent trends (Blackport and Fyfe 2022; Eade et al. 2022) and also in the seasonal predictions based on these hindcasts (Dunstone et al. 2016, 2023). NAO-like circulation patterns are already seen to contribute to the spread in SAT trends, as seen in Fig. 7.

Figure 9a shows trends in the winter NAO anomaly index calculated following Dunstone et al. (2016). The winter NAO index has a weak positive trend between 1981 and 2022 ($p = 0.76$ in ERA5), falling within the hindcast range at both lead times and even within the five-member uninitialized ensemble. The spread

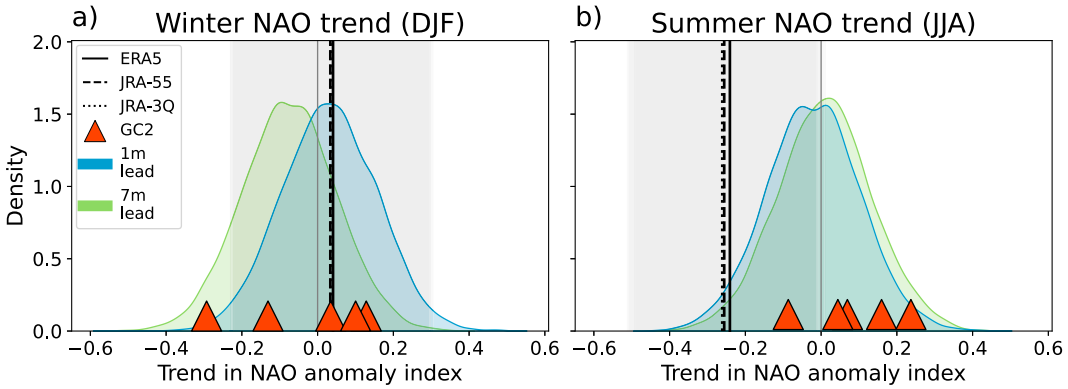

FIG. 9. Trends in (a) the DJF winter NAO anomaly index (calculated following Dunstone et al. 2016) and (b) the JJA SNAO anomaly index (calculated following Dunstone et al. 2016). Colors and symbols are as in Fig. 5. Trends are calculated over 1981/82–2022/23 for DJF and 1981–2022 for JJA.

of hindcast trends is large, such that trends of either sign are clearly possible within the ensemble. Previous work has drawn attention to the winter NAO trend over the longer period since 1950, which is at the upper edge of the trends simulated by free-running CMIP6 models (Blackport and Fyfe 2022), and includes a particularly large positive trend between the 1960s and 1990s (Eade et al. 2022). As our trends begin in 1981/82, we are unable to test the model's ability to capture trends beginning in the mid-twentieth century. Over the more recent period considered here, we find that both the uninitialized model and the hindcasts are capable of capturing the modest winter NAO trend.

A much stronger reanalysis trend is seen in the summer NAO (SNAO; Fig. 9b). Here, the 95% confidence intervals on the reanalysis regressions all lie entirely below zero, and the regressions achieve higher statistical significance than in winter (e.g., $p = 0.07$ in ERA5). Fewer than 4% of hindcast members at 1-m lead and 3% of hindcast members at 7-m lead have as strong a negative trend as the reanalysis. Interestingly, the hindcast SNAO trend distribution is largely insensitive to lead time. This may reflect the poor seasonal predictability of the European summer climate (Patterson et al. 2022). Recent work has also pointed to the importance of aerosol forcing in summer circulation trends over the European sector (Dong and Sutton 2021); aerosol forcing is identical in the hindcasts at both lead times and in the free-running GC2 simulations, which may explain their similar trend distributions.

In summary, the hindcasts support a wide range of positive and negative trends in the NAO indices in both seasons. While long-term trends in the winter NAO have received considerable attention in the literature, we find that both the hindcasts and the free-running GC2 model are able to capture observed trends between 1981 and 2022. However, the trend toward negative SNAO over the same period is further toward the edge of the hindcast trend distributions, which may point to model deficiencies.

7. Discussion

The hindcast trends in previous sections are constructed by randomly sampling individual ensemble members in each year, with each segment of the time series sampled independently of the previous segments. To determine whether this approach yields physically plausible time series, we perform statistical tests to assess the consistency between the bootstrapped hindcast time series and the observed climate. Following Kelder et al. (2020), we calculate the mean, standard deviation, skewness, and kurtosis of the hindcast time series and test whether the corresponding reanalysis values fall within their 5%–95% confidence intervals. The test is performed for a selection of the quantitative metrics presented above: jet indices, NAO indices, and GSAT. Since the trends sample year-to-year variability in the hindcasts, consistency in the standard deviation, skewness, and kurtosis builds confidence that the spread of trends is physically plausible.

Figure S3 shows the statistical moments for the jet indices used in section 4. The hindcasts show mean biases of varying sizes relative to the reanalyses in each season and hemisphere. Despite this, the higher statistical moments in the reanalyses are all within the 5%–95% confidence intervals for both the 1-m lead and 7-m lead hindcasts, with the exception of the kurtosis in DJF NH. Likewise, the statistical moments of the NAO and SNAO distributions are well captured by the hindcasts at both lead times (Fig. S4). The hindcast jet and NAO trends thus sample variability that is consistent with the observed climate.

The statistical moments for the GSAT time series are shown in Fig. S5. The hindcasts are biased cold relative to the reanalyses in both seasons. Unlike for the jet indices and NAO, however, the higher moments of the reanalysis GSAT time series are not consistently well captured. Focusing on the 1-m hindcasts, none of the reanalysis products fall within the 5%–95% confidence intervals in DJF, and in JJA, only one product falls within the interval for the standard deviation and none for the kurtosis. This suggests that caution is necessary when interpreting the spread in hindcast GSAT trends. However, we also note that the spread between the reanalyses is much larger for GSAT than for the jet and NAO indices; for example, the DJF GSAT standard deviation in ERA5 is 20% larger than the average of the JRA products (17% larger in JJA). In both seasons, the standard deviation values of the

1-m hindcasts are intermediate between the ERA5 and JRA values. Hence, while the hindcast GSAT time series do not achieve strict statistical consistency with the reanalyses, the poor agreement between the different reanalysis products limits our ability to conclude that the hindcast trends are physically implausible.

8. Conclusions

This study has outlined a method for exploring the full range of multidecadal trends in an ensemble of interannual hindcasts (DePreSys3; Dunstone et al. 2016). Part of the motivation was to explore where the spread of hindcast trends is large and cannot be neglected and where the conventional ensemble-mean approach may be appropriate. The main results are as follows:

- The 1-m hindcasts broadly support the emergence of poleward zonal-mean jet shifts on the global scale (Woollings et al. 2023), with 85% of members showing concurrent poleward shifts in both hemispheres in DJF and 96% in JJA. However, the magnitude of the shifts varies substantially across the ensemble, even in hemispheres and seasons where the trend is thought to be strongly forced; for example, between 8% and 11% of hindcast members show an equatorward shift of the austral summer jet, despite shared ozone forcing and a strong poleward shift in the ensemble mean.
- The spread in 1-m hindcast GSAT trends is 0.03 K decade^{-1}, equivalent to approximately a sixth of the total trend over 1981–2022 or a tenth of the observed warming since preindustrial times (Gulev et al. 2021). In DJF, 87% of the spread in GSAT trends occurs over land, which can change temperature quickly in response to the advection of mild oceanic air by the atmospheric circulation (Wallace et al. 1995). Comparisons to observed SAT trends, particularly for trends over extratropical land regions, should therefore use the full range of hindcast trends where possible.
- Variability in the hindcast time series is statistically consistent with the reanalyses for the jet and NAO indices, building confidence in the physical plausibility of the hindcast trends. The hindcast GSAT time series are statistically distinct from the reanalyses; however, while we should be cautious in interpreting the spread of GSAT trends, we also note substantial uncertainty in the statistical moments of the reanalyses.
- Seasonal variability leads to a wide spread in tropospheric temperature trends at high latitudes, particularly in the Arctic winter. By contrast, the range of tropical temperature trends in the hindcasts is much smaller, with smaller deviations from the ensemble-mean trend.

The overarching theme of our work has been to demonstrate the potential for short-time-scale variability to contribute to multidecadal trends in a variety of climate metrics. Variability on seasonal time scales is likely to be mostly atmospheric in origin, although fast coupled processes may also contribute. Accordingly, the spread of trends is particularly large for trends in atmospheric circulation features (e.g., the zonal-mean jets and NAO, Figs. 5 and 9), consistent with the low forced signal of the circulation relative to internal noise (Shepherd 2014). However, the dynamics also drive a nonnegligible spread in temperature trends, particularly at higher latitudes, on regional scales, and over land (Figs. 3, 6, and 7). The importance of internal atmospheric variability for multidecadal trends is well established in studies of climate change (e.g., Deser et al. 2012; Jain et al. 2023), but it appears to have been less emphasized for understanding trends in seasonal forecast models; to our knowledge, ours is the first study that has taken this approach. This study demonstrates that accounting for the spread of trends is both possible and necessary when comparing trends in forecast models to observations, with implications for studies that have drawn conclusions based on ensemble-mean trend errors.

More broadly, hindcasts offer a largely untapped resource for studying the influence of short-time-scale variability on long-term climate trends. The contribution of atmospheric variability to multidecadal trends has typically been studied using atmospheric models forced by prescribed SSTs. However, the lack of physical coupling in atmosphere-only models can lead to underestimates of low-frequency extratropical variability (Barsugli and Battisti 1998; Mori et al. 2024), and the oceanic influence on the atmosphere is not always directly mediated by SST anomalies (Sutton and Mathieu 2002). Generating large model ensembles also remains computationally expensive, even in an uncoupled configuration. We have shown that a very large model ensemble of plausible multidecadal trends can be generated using existing coupled hindcast data, of the type that is routinely archived by operational centers. The hindcast trend ensemble offers a complementary view to prescribed SST ensembles, allowing for some faster coupled processes in addition to purely atmospheric variability. The very large ensemble size also means that the conditional subsampling of hindcast members (as in Fig. 7), which is useful for building physical understanding of mechanisms, can yield robust statistics even for rare manifestations of internal variability that appear in the tails of the trend distribution. The hindcast trend ensemble introduced here may therefore provide a useful tool for contextualizing emerging trends in the climate system.

Acknowledgments. The authors thank Mika Rantanen and two anonymous reviewers for their thorough and helpful comments. R. T. was funded by the Natural Environment Research Council (NERC), Doctoral Training Partnership in Environmental Research (NE/S007474/1). N. D. was supported by the Met Office Hadley Centre Climate Programme funded by BEIS and Defra. Analysis was performed on JASMIN, operated by the Science and Technology Facilities Council (STFC) on behalf of NERC.

Data availability statement. ERA5 reanalysis data are available from the Copernicus Climate Change Service (C3S) at Climate Data Store (CDS; https://cds.climate.copernicus.eu/), operated by the European Centre for Medium-Range Weather Forecasts (ECMWF). DePreSys3 and HadGEM3-GC2 data are available upon reasonable request from the authors.

REFERENCES

Banerjee, A., J. C. Fyfe, L. M. Polvani, D. Waugh, and K.-L. Chang, 2020: A pause in Southern Hemisphere circulation trends due to the Montreal Protocol. *Nature*, **579**, 544–548, https://doi.org/10.1038/s41586-020-2120-4.

Barsugli, J. J., and D. S. Battisti, 1998: The basic effects of atmosphere–ocean thermal coupling on midlatitude variability. *J. Atmos. Sci.*, **55**, 477–493, https://doi.org/10.1175/1520-0469(1998)055<0477:TBEOAO>2.0.CO;2.

Becker, E. J., B. P. Kirtman, M. L'Heureux, A. G. Munoz, and K. Pegion, 2022: A decade of the North American Multimodel Ensemble (NMME): Research, application, and future directions. *Bull. Amer. Meteor. Soc.*, **103**, E973–E995, https://doi.org/10.1175/BAMS-D-20-0327.1.

Bengtsson, L., S. Hagemann, and K. I. Hodges, 2004: Can climate trends be calculated from reanalysis data? *J. Geophys. Res.*, **109**, D11111, https://doi.org/10.1029/2004JD004536.

Best, M. J., and Coauthors, 2011: The Joint UK Land Environment Simulator (JULES), model description—Part 1: Energy and water fluxes. *Geosci. Model Dev.*, **4**, 677–699, https://doi.org/10.5194/gmd-4-677-2011.

Beverley, J. D., M. Newman, and A. Hoell, 2024: Climate model trend errors are evident in seasonal forecasts at short leads. *npj Climate Atmos. Sci.*, **7**, 285, https://doi.org/10.1038/s41612-024-00832-w.

Blackport, R., and J. C. Fyfe, 2022: Climate models fail to capture strengthening wintertime North Atlantic jet and impacts on Europe. *Sci. Adv.*, **8**, eabn3112, https://doi.org/10.1126/sciadv.abn3112.

Bowler, N. E., A. Arribas, S. E. Beare, K. R. Mylne, and G. J. Shutts, 2009: The local ETKF and SKEB: Upgrades to the MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **135**, 767–776, https://doi.org/10.1002/qj.394.

Choi, J., S.-W. Son, Y.-G. Ham, J.-Y. Lee, and H.-M. Kim, 2016: Seasonal-to-interannual prediction skills of near-surface air temperature in the CMIP5 decadal hindcast experiments. *J. Climate*, **29**, 1511–1527, https://doi.org/10.1175/JCLI-D-15-0182.1.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

Deser, C., A. Phillips, V. Bourdette, and H. Teng, 2012: Uncertainty in climate change projections: The role of internal variability. *Climate Dyn.*, **38**, 527–546, https://doi.org/10.1007/s00382-010-0977-x.

Dong, B., and R. T. Sutton, 2021: Recent trends in summer atmospheric circulation in the North Atlantic/European region: Is there a role for anthropogenic aerosols? *J. Climate*, **34**, 6777–6795, https://doi.org/10.1175/JCLI-D-20-0665.1.

Dunstone, N., D. Smith, A. Scaife, L. Hermanson, R. Eade, N. Robinson, M. Andrews, and J. Knight, 2016: Skilful predictions of the winter North Atlantic Oscillation one year ahead. *Nat. Geosci.*, **9**, 809–814, https://doi.org/10.1038/ngeo2824.

——, and Coauthors, 2023: Skilful predictions of the summer North Atlantic Oscillation. *Commun. Earth Environ.*, **4**, 409, https://doi.org/10.1038/s43247-023-01063-2.

Eade, R., D. B. Stephenson, A. A. Scaife, and D. M. Smith, 2022: Quantifying the rarity of extreme multi-decadal trends: How unusual was the late twentieth century trend in the North Atlantic Oscillation? *Climate Dyn.*, **58**, 1555–1568, https://doi.org/10.1007/s00382-021-05978-4.

Gillett, N. P., and D. W. J. Thompson, 2003: Simulation of recent Southern Hemisphere climate change. *Science*, **302**, 273–275, https://doi.org/10.1126/science.1087440.

Gulev, S., and Coauthors, 2021: Changing state of the climate system. *Climate Change 2021: The Physical Science Basis*, V. Masson-Delmotte et al., Eds., Cambridge University Press, 287–422.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Iles, C., and G. Hegerl, 2017: Role of the North Atlantic Oscillation in decadal temperature trends. *Environ. Res. Lett.*, **12**, 114010, https://doi.org/10.1088/1748-9326/aa9152.

Jain, S., A. A. Scaife, T. G. Shepherd, C. Deser, N. Dunstone, G. A. Schmidt, K. E. Trenberth, and T. Turkington, 2023: Importance of internal variability for climate model assessment. *npj Climate Atmos. Sci.*, **6**, 68, https://doi.org/10.1038/s41612-023-00389-0.

Kay, G., N. J. Dunstone, D. M. Smith, R. A. Betts, C. Cunningham, and A. A. Scaife, 2022: Assessing the chance of unprecedented dry conditions over North Brazil during El Niño events. *Environ. Res. Lett.*, **17**, 064016, https://doi.org/10.1088/1748-9326/ac6df9.

Kelder, T., and Coauthors, 2020: Using UNSEEN trends to detect decadal changes in 100-year precipitation extremes. *npj Climate Atmos. Sci.*, **3**, 47, https://doi.org/10.1038/s41612-020-00149-4.

Kobayashi, S., and Coauthors, 2015: The JRA-55 reanalysis: General specifications and basic characteristics. *J. Meteor. Soc. Japan*, **93**, 5–48, https://doi.org/10.2151/jmsj.2015-001.

Kosaka, Y., and Coauthors, 2024: The JRA-3Q reanalysis. *J. Meteor. Soc. Japan*, **102**, 49–109, https://doi.org/10.2151/jmsj.2024-004.

Krakauer, N. Y., 2019: Temperature trends and prediction skill in NMME seasonal forecasts. *Climate Dyn.*, **53**, 7201–7213, https://doi.org/10.1007/s00382-017-3657-2.

Kushnir, Y., and Coauthors, 2019: Towards operational predictions of the near-term climate. *Nat. Climate Change*, **9**, 94–101, https://doi.org/10.1038/s41558-018-0359-7.

L'Heureux, M. L., M. K. Tippett, and W. Wang, 2022: Prediction challenges from errors in tropical Pacific sea surface temperature trends. *Front. Climate*, **4**, 837483, https://doi.org/10.3389/fclim.2022.837483.

Megann, A., and Coauthors, 2014: GO5.0: The joint NERC–Met Office NEMO global ocean model for use in coupled and forced applications. *Geosci. Model Dev.*, **7**, 1069–1092, https://doi.org/10.5194/gmd-7-1069-2014.

Mitchell, D. M., Y. T. E. Lo, W. J. M. Seviour, L. Haimberger, and L. M. Polvani, 2020: The vertical profile of recent tropical temperature trends: Persistent model biases in the context of internal variability. *Environ. Res. Lett.*, **15**, 1040b4, https://doi.org/10.1088/1748-9326/ab9af7.

Mori, M., Y. Kosaka, B. Taguchi, H. Tokinaga, H. Tatebe, and H. Nakamura, 2024: Northern Hemisphere winter atmospheric teleconnections are intensified by extratropical ocean-atmosphere coupling. *Commun. Earth Environ.*, **5**, 124, https://doi.org/10.1038/s43247-024-01282-1.

Patterson, M., A. Weisheimer, D. J. Befort, and C. H. O'Reilly, 2022: The strong role of external forcing in seasonal forecasts of European summer temperature. *Environ. Res. Lett.*, **17**, 104033, https://doi.org/10.1088/1748-9326/ac9243.

Rae, J. G. L., H. T. Hewitt, A. B. Keen, J. K. Ridley, A. E. West, C. M. Harris, E. C. Hunke, and D. N. Walters, 2015:

Development of the global sea ice 6.0 CICE configuration for the Met Office Global Coupled model. *Geosci. Model Dev.*, **8**, 2221–2230, https://doi.org/10.5194/gmd-8-2221-2015.

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. Geophys. Res.*, **108**, 4407, https://doi.org/10.1029/2002JD002670.

Santer, B. D., T. M. L. Wigley, J. S. Boyle, D. J. Gaffen, J. J. Hnilo, D. Nychka, D. E. Parker, and K. E. Taylor, 2000: Statistical significance of trends and trend differences in layer-average atmospheric temperature time series. *J. Geophys. Res.*, **105**, 7337–7356, https://doi.org/10.1029/1999JD901105.

Senior, C. A., and Coauthors, 2016: Idealized climate change simulations with a high-resolution physical model: HadGEM3-GC2. *J. Adv. Model. Earth Syst.*, **8**, 813–830, https://doi.org/10.1002/2015MS000614.

Seviour, W. J. M., D. W. Waugh, L. M. Polvani, G. J. P. Correa, and C. I. Garfinkel, 2017: Robustness of the simulated tropospheric response to ozone depletion. *J. Climate*, **30**, 2577–2585, https://doi.org/10.1175/JCLI-D-16-0817.1.

Shao, Y., Q. J. Wang, A. Schepen, and D. Ryu, 2021: Going with the trend: Forecasting seasonal climate conditions under climate change. *Mon. Wea. Rev.*, **149**, 2513–2522, https://doi.org/10.1175/MWR-D-20-0318.1.

Shepherd, T. G., 2014: Atmospheric circulation as a source of uncertainty in climate change projections. *Nat. Geosci.*, **7**, 703–708, https://doi.org/10.1038/ngeo2253.

Shin, C.-S., and B. Huang, 2019: A spurious warming trend in the NMME equatorial Pacific SST hindcasts. *Climate Dyn.*, **53**, 7287–7303, https://doi.org/10.1007/s00382-017-3777-8.

Simmons, A. J., 2022: Trends in the tropospheric general circulation from 1979 to 2022. *Wea. Climate Dyn.*, **3**, 777–809, https://doi.org/10.5194/wcd-3-777-2022.

——, and Coauthors, 2004: Comparison of trends and low-frequency variability in CRU, ERA-40, and NCEP/NCAR analyses of surface air temperature. *J. Geophys. Res.*, **109**, D24115, https://doi.org/10.1029/2004JD005306.

Simpson, I. R., K. A. McKinnon, F. V. Davenport, M. Tingley, F. Lehner, A. A. Fahad, and D. Chen, 2021: Emergent constraints on the large-scale atmospheric circulation and regional hydroclimate: Do they still work in CMIP6 and how much can they actually constrain the future? *J. Climate*, **34**, 6355–6377, https://doi.org/10.1175/JCLI-D-21-0055.1.

Smith, D. M., and Coauthors, 2015: Earth's energy imbalance since 1960 in observations and CMIP5 models. *Geophys. Res. Lett.*, **42**, 1205–1213, https://doi.org/10.1002/2014GL062669.

Son, S.-W., and Coauthors, 2010: Impact of stratospheric ozone on Southern Hemisphere circulation change: A multimodel assessment. *J. Geophys. Res.*, **115**, D00M07, https://doi.org/10.1029/2010JD014271.

Sutton, R., and P.-P. Mathieu, 2002: Response of the atmosphere–ocean mixed-layer system to anomalous ocean heat-flux convergence. *Quart. J. Roy. Meteor. Soc.*, **128**, 1259–1275, https://doi.org/10.1256/003590002320373283.

Trenberth, K. E., 1997: The definition of El Niño. *Bull. Amer. Meteor. Soc.*, **78**, 2771–2778, https://doi.org/10.1175/1520-0477(1997)078<2771:TDOENO>2.0.CO;2.

——, and D. J. Shea, 2006: Atlantic hurricanes and natural variability in 2005. *Geophys. Res. Lett.*, **33**, L12704, https://doi.org/10.1029/2006GL026894.

Wallace, J. M., Y. Zhang, and J. A. Renwick, 1995: Dynamic contribution to hemispheric mean temperature trends. *Science*, **270**, 780–783, https://doi.org/10.1126/science.270.5237.780.

——, ——, and L. Bajuk, 1996: Interpretation of interdecadal trends in Northern Hemisphere surface air temperature. *J. Climate*, **9**, 249–259, https://doi.org/10.1175/1520-0442(1996)009<0249:IOITIN>2.0.CO;2.

Wilks, D. S., 2016: "The stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. *Bull. Amer. Meteor. Soc.*, **97**, 2263–2273, https://doi.org/10.1175/BAMS-D-15-00267.1.

Williams, K. D., and Coauthors, 2015: The Met Office Global Coupled model 2.0 (GC2) configuration. *Geosci. Model Dev.*, **8**, 1509–1524, https://doi.org/10.5194/gmd-8-1509-2015.

Woollings, T., M. Drouard, C. H. O'Reilly, D. M. H. Sexton, and C. McSweeney, 2023: Trends in the atmospheric jet streams are emerging in observations and could be linked to tropical warming. *Commun. Earth Environ.*, **4**, 125, https://doi.org/10.1038/s43247-023-00792-8.