Supplement of Geosci. Model Dev., 14, 7073–7116, 2021
https://doi.org/10.5194/gmd-14-7073-2021-supplement

Geoscientific Model Development
Open Access

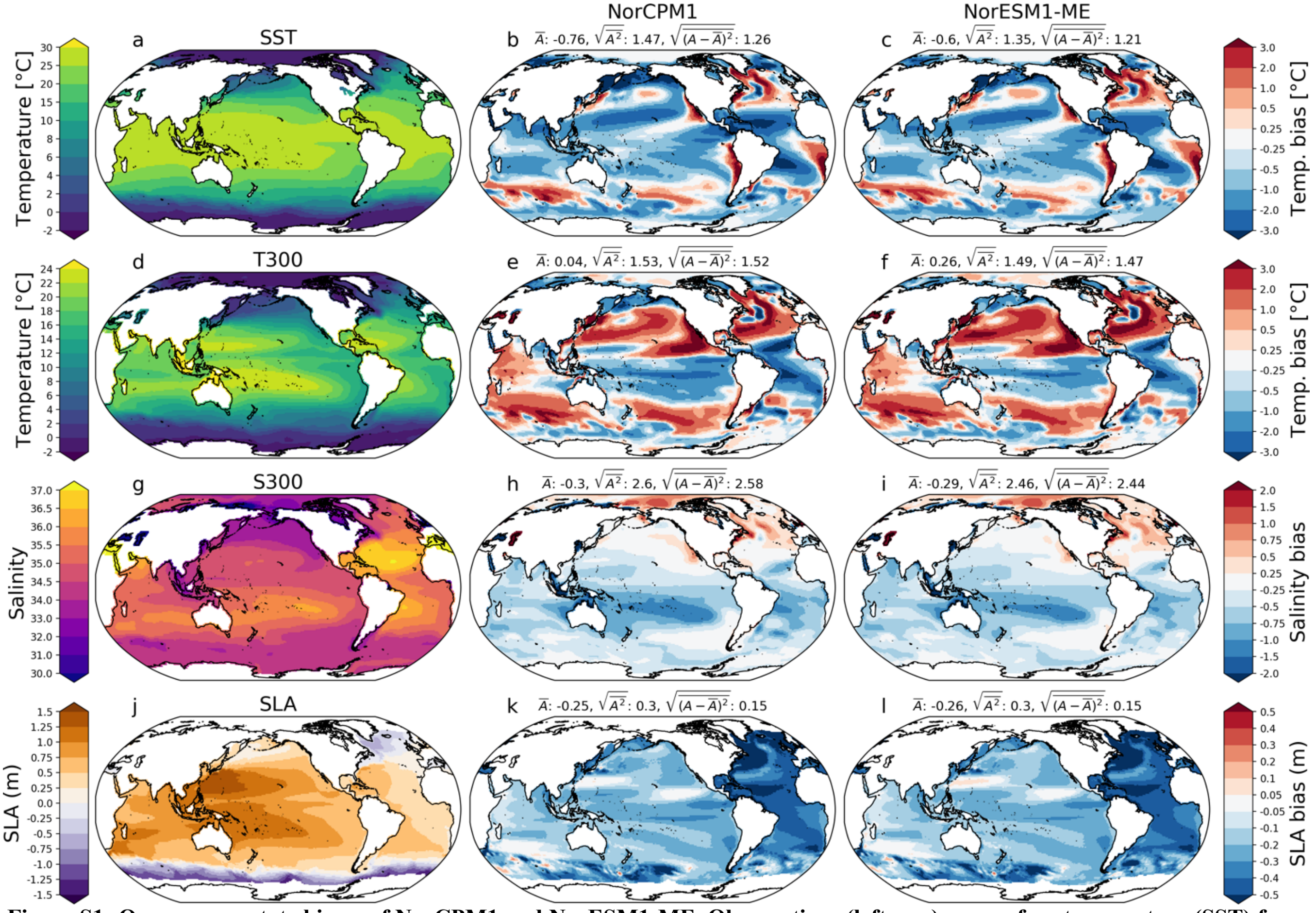
Supplement of

NorCPM1 and its contribution to CMIP6 DCPP

Ingo Bethke et al.

Correspondence to: Ingo Bethke (ingo.bethke@uib.no)

S1 Earth system model evaluation



NorCPM1
NorESM1-ME
a
SST
b
$\bar{A}$: -0.76, $\sqrt{\bar{A^2}}$: 1.47, $\sqrt{\overline{(A-\bar{A})^2}}$: 1.26
c
$\bar{A}$: -0.6, $\sqrt{\bar{A^2}}$: 1.35, $\sqrt{\overline{(A-\bar{A})^2}}$: 1.21
d
T300
e
$\bar{A}$: 0.04, $\sqrt{\bar{A^2}}$: 1.53, $\sqrt{\overline{(A-\bar{A})^2}}$: 1.52
f
$\bar{A}$: 0.26, $\sqrt{\bar{A^2}}$: 1.49, $\sqrt{\overline{(A-\bar{A})^2}}$: 1.47
g
S300
h
$\bar{A}$: -0.3, $\sqrt{\bar{A^2}}$: 2.6, $\sqrt{\overline{(A-\bar{A})^2}}$: 2.58
i
$\bar{A}$: -0.29, $\sqrt{\bar{A^2}}$: 2.46, $\sqrt{\overline{(A-\bar{A})^2}}$: 2.44
j
SLA
k
$\bar{A}$: -0.25, $\sqrt{\bar{A^2}}$: 0.3, $\sqrt{\overline{(A-\bar{A})^2}}$: 0.15
l
$\bar{A}$: -0.26, $\sqrt{\bar{A^2}}$: 0.3, $\sqrt{\overline{(A-\bar{A})^2}}$: 0.15
Temperature [°C]
30 25 20 15 10 8 6 4 2 0 -2
Temperature [°C]
24 22 20 18 16 14 12 10 8 6 4 2 0 -2
Salinity
37.0 36.5 36.0 35.5 35.0 34.5 34.0 33.0 32.0 31.0 30.0
SLA (m)
1.5 1.25 1.0 0.75 0.5 0.25 0.0 -0.25 -0.5 -0.75 -1.0 -1.25 -1.5
Temp. bias [°C]
3.0 2.0 1.0 0.5 0.25 -0.25 -0.5 -1.0 -2.0 -3.0
Temp. bias [°C]
3.0 2.0 1.0 0.5 0.25 -0.25 -0.5 -1.0 -2.0 -3.0
Salinity bias
2.0 1.5 1.0 0.75 0.5 0.25 -0.25 -0.5 -0.75 -1.0 -1.5 -2.0
SLA bias (m)
0.5 0.4 0.3 0.2 0.1 0.05 -0.05 -0.1 -0.2 -0.3 -0.4 -0.5

Figure S1: Ocean mean-state biases of NorCPM1 and NorESM1-ME. Observations (left row) use surface temperature (SST) from 1971–2000 (Reyn_SmithOIv2; Reynolds et al., 2002), 0–300 m temperature (T300), and salinity (S300) from 1980–2010 (EN4.2.1; Good et al., 2013), and sea level anomaly (SLA) from 1993–2018 (ARMOR3D; Larnicol et al., 2006). Matching time periods are used for computing the climatological biases of NorCPM1 (middle row) and NorESM1-ME (right row). Area-weighted global-mean bias, RMSE, and global-mean bias adjusted RMSE are stated above the panels.

In this section, we evaluated climate mean and variability characteristics, model stability and sensitivity to external forcings of the ESM component of NorCPM1 in absence of data assimilation. Knowledge of the model climate performance can inform attribution of inter-model prediction differences and generally the interpretation of prediction results. The mean climate and variability of the original NorESM1 has been evaluated in previous studies (e.g., Bentsen et al., 2013; Iversen et al., 2013; Tjiputra et al., 2013; Kirkevåg et al., 2013). The updates to the code and forcings have not changed the model characteristics to an extent that would warrant a complete re-evaluation and the evaluation here is therefore kept brief.

Figures S1 and S2 show the climatological biases of NorCPM1 and NorESM1-ME for the ocean and atmosphere, respectively, based on the late 20th Century state of *historical*. NorCPM1's simulated sea surface temperature (SST) is 0.7 K too cold globally, with larger spatial biases (Fig. S1b). The tropical Pacific is slightly too cold, and the subtropics and polar regions are generally too cold by several K. In turn, the Southern Ocean, the subpolar North Atlantic (SPNA) except south of Greenland, coastal upwelling regions, and western boundary current extensions are generally too warm. Upper ocean temperature (T300, 0-300 m averaged) shows a similar pattern (Fig. S1e), indicating that the surface biases largely extend below the surface. For parts of the central North Pacific and Indian Ocean, however, the sign of the T300 biases is opposite compared to the SST biases, indicating a contribution from errors in vertical mixing. The warm bias in the SPNA is accompanied by a negative sea surface height (SSH) bias (Fig. S1k) in excess of 1 m and a positive upper ocean salt (S300, 0-300 m averaged) bias (Fig. S1h) that both extend into the Arctic. This salt bias is mirrored by a fresh bias south of the equator, suggesting large-scale ocean meridional circulation errors as one cause. The simulated Atlantic Meridional Overturning Circulation (AMOC) is indeed too vigorous (Fig. S3c), with a mean strength of 34 Sv versus 19 Sv observational estimate (Cunningham et al., 2007) evaluated at 26.5 °N, and the northward oceanic heat transport in the Atlantic Ocean is too large (Fig. S3b). It should be noted, however, that a similar configuration of NorESM1 featuring a weaker AMOC produces similar temperature and heat transport biases (Guo et al., 2019). Surface air temperature (SAT, evaluated at 2 m) biases (Fig. S2b) match those of SST except over ice-covered regions where the model simulates cold biases in excess of 5 K. The Arctic cold bias extends over much of the North American and Eurasian continents, including North Africa and East Asia. Warm biases are simulated over land adjacent to upwelling coastal upwelling regions and over central Asia, whereas only small temperature biases are simulated over most of Europe and central and eastern part of South America. The Arctic cold bias is accompanied by a negative 500 hPa geopotential height (Z500) bias that is strongest over the Chukchi Sea and Nordic Seas and opposed by positive Z500 biases over the North Pacific and subtropical North Atlantic (Fig. S2k), resulting in a too zonal storm track over the North Atlantic. Similarly, cold surface biases over Antarctica and warm biases over the Southern Ocean are accompanied by negative and positive Z500 biases. The precipitation biases over the ocean reveal the presence of a double inter-tropical convergence zone (Fig. S2h). Over land (Fig. S2k), wet biases are present over much of Africa, the Tibetan plateau, Australia, and westerns parts of America, while dry biases are present over the remainder of America, North Africa, India and Northern Eurasia.

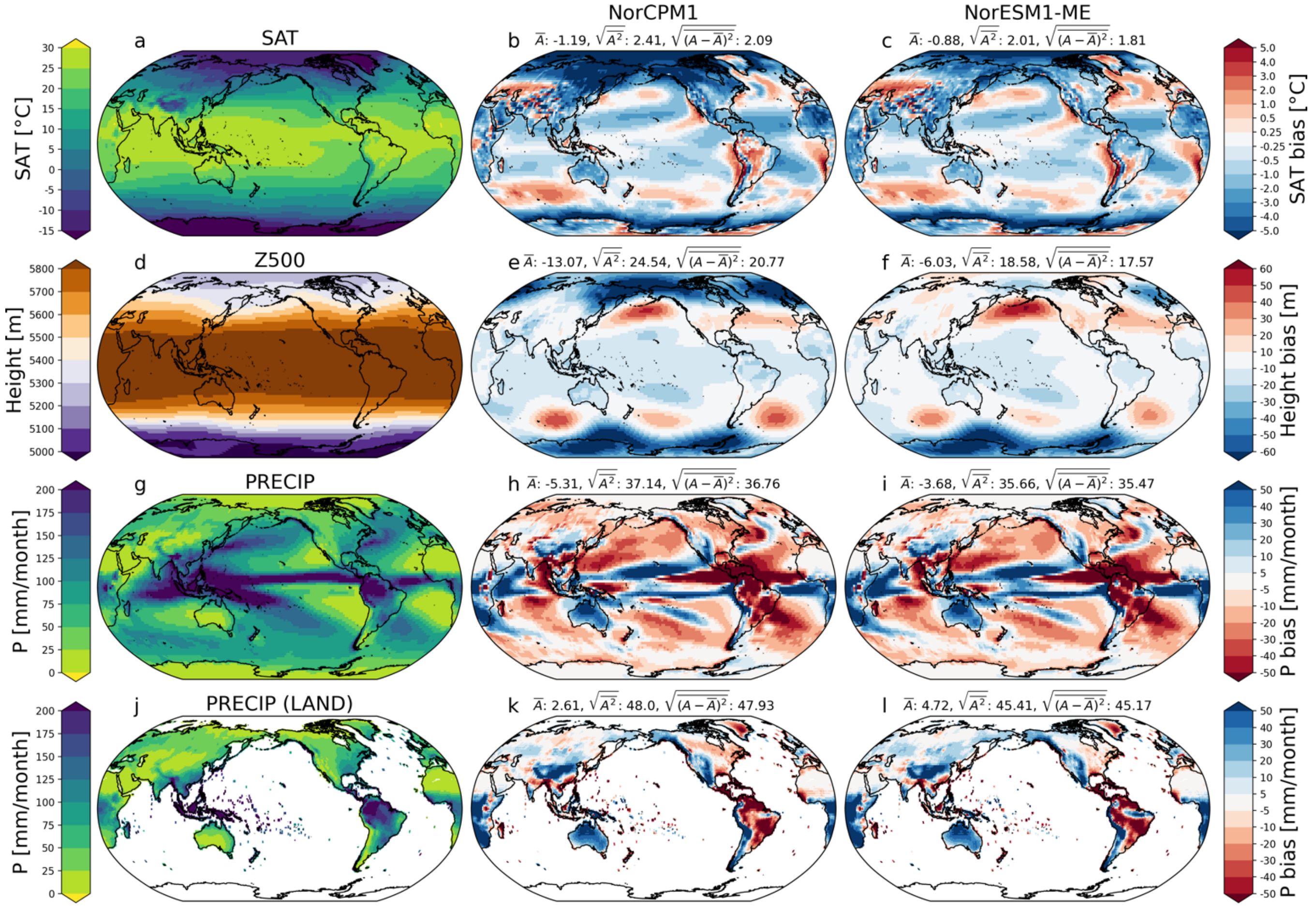

NorCPM1
NorESM1-ME
a
SAT
b $\bar{A}$: -1.19, $\sqrt{\overline{A^2}}$: 2.41, $\sqrt{\overline{(A-\bar{A})^2}}$: 2.09
c $\bar{A}$: -0.88, $\sqrt{\overline{A^2}}$: 2.01, $\sqrt{\overline{(A-\bar{A})^2}}$: 1.81
SAT [°C]
30 25 20 15 10 5 0 -5 -10 -15
SAT bias [°C]
5.0 4.0 3.0 2.0 1.0 0.5 0.25 -0.25 -0.5 -1.0 -2.0 -3.0 -4.0 -5.0
d
Z500
e $\bar{A}$: -13.07, $\sqrt{\overline{A^2}}$: 24.54, $\sqrt{\overline{(A-\bar{A})^2}}$: 20.77
f $\bar{A}$: -6.03, $\sqrt{\overline{A^2}}$: 18.58, $\sqrt{\overline{(A-\bar{A})^2}}$: 17.57
Height [m]
5800 5700 5600 5500 5400 5300 5200 5100 5000
Height bias [m]
60 50 40 30 20 10 -10 -20 -30 -40 -50 -60
g
PRECIP
h $\bar{A}$: -5.31, $\sqrt{\overline{A^2}}$: 37.14, $\sqrt{\overline{(A-\bar{A})^2}}$: 36.76
i $\bar{A}$: -3.68, $\sqrt{\overline{A^2}}$: 35.66, $\sqrt{\overline{(A-\bar{A})^2}}$: 35.47
P [mm/month]
200 175 150 125 100 75 50 25 0
P bias [mm/month]
50 40 30 20 10 5 -5 -10 -20 -30 -40 -50
j
PRECIP (LAND)
k $\bar{A}$: 2.61, $\sqrt{\overline{A^2}}$: 48.0, $\sqrt{\overline{(A-\bar{A})^2}}$: 47.93
l $\bar{A}$: 4.72, $\sqrt{\overline{A^2}}$: 45.41, $\sqrt{\overline{(A-\bar{A})^2}}$: 45.17
P [mm/month]
200 175 150 125 100 75 50 25 0
P bias [mm/month]
50 40 30 20 10 5 -5 -10 -20 -30 -40 -50

Figure S2: Atmosphere mean-state biases of NorCPM1 and NorESM1-ME. Observations (left row) use 2 m air temperature (SAT), geopotential height at 500 hPa (Z500), and global precipitation from 1979–2019 (ERA5; Hersbach et al., 2020), and precipitation over land from 1979–2018 (CRU TS v4.03; Harris et al., 2020). Matching time periods are used for computing the climatological biases of NorCPM1 (middle row) and NorESM1-ME (right row). Area-weighted global-mean bias, RMSE, and global-mean bias adjusted RMSE are stated above the panels.

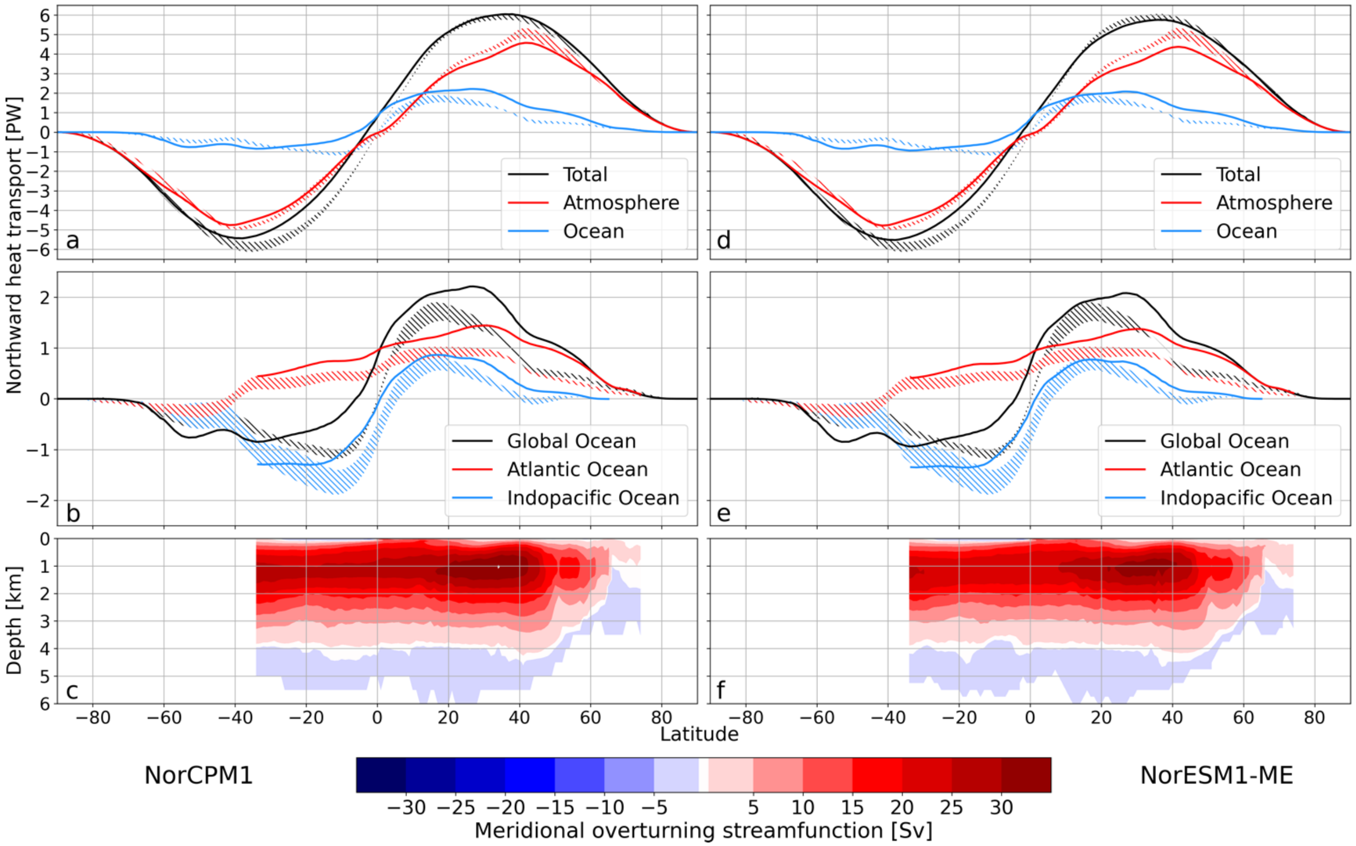


Figure S3: Atlantic meridional transports of heat and mass. Simulated historical (1976–2005 mean) northward heat transport of NorCPM1 for (a) global atmosphere, ocean, and total, and (b) global ocean, its decomposed Atlantic Ocean and Pacific and Indian oceans, and Atlantic meridional overturning circulation (AMOC) streamfunction (c). The panels (d–f) show the same for NorESM1-ME. The corresponding hatched areas with uncertainties are estimates from Fasullo and Trenberth (2008). In the model estimation, the ocean heat transport is calculated directly from the ocean model, and the atmospheric heat transport is derived by meridional integration of the difference between zonally integrated net TOA and surface heat fluxes.

While the climate mean-state biases are overall similar to those presented in the original NorESM1 description paper (Bentsen et al., 2013), Figures S1–S3 reveal signs of degradation for NorCPM1, with global mean biases and RMSEs being slightly larger than for NorESM1-ME. Most notable, the Arctic cold and negative Z500 biases are exacerbated. The bias exacerbation is likely due to erroneous transient land-cryosphere surface conditions in NorCPM1, as revealed from additional sensitivity simulations. Each sensitivity simulation has one NorCPM1's forcing update or model modification reverted back to the NorESM1-ME CMIP5 system; these were (1) aerosol code bug, (2) atmospheric cloud tuning, (3) sea ice albedo tuning, (4) land surface forcing, (5) ozone forcing and (6) stratospheric volcanic forcing. We integrated the simulations for 120 years with 1850-level preindustrial forcings, followed by 170 years with transient historical forcings. We evaluate the preindustrial (year-1850) state by averaging the first 120 simulation years and the modern (year-2000) state by averaging the period corresponding to the forcing years 1879-2019. Figure S4 shows the results for SAT and Z500 averaged over the Arctic region north of 60 °N. Only the simulation with land-use change reverted back to CMIP5 (purple lines) shows a SAT warming trend and Z500 rise comparable to the NorESM1-ME CMIP5 system. The impacts of the other changes are comparably small—the aerosol code bug leads to some high-latitude warming, whereas reverting to CMIP5 ozone forcing or sea ice tuning leads to some cooling—and do not significantly affect the trend. Rectifying NorCPM1's land surface conditions will be a future development priority. A comparison with results from previous NorCPM versions suggests that the increased biases have a measurable detrimental effect on high latitude prediction skill (Passos et al, under review).

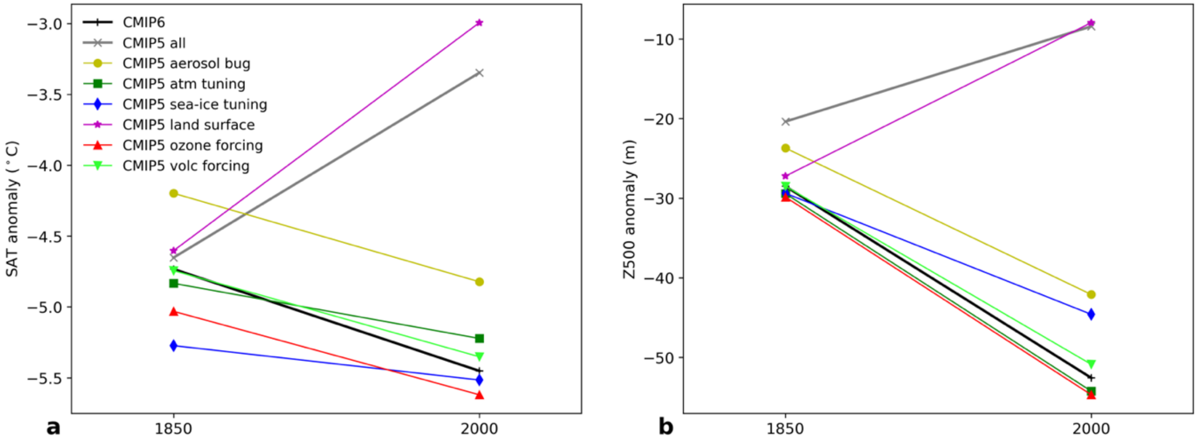


Figure S4: Arctic SAT (a) and Z500 (b) biases for preindustrial and modern time from supporting sensitivity experiments (see text for details). The sensitivity experiments (colours) are based on NorCPM1 (black), but with individual updates reverted back to NorESM1-ME (grey). Note the differences between a sensitivity experiment to NorCPM1 is opposite to the effect of the specific upgrade (e.g., the effect of the aerosol bug fix is a cooling, not warming).

Global mean surface temperature (GMST) is commonly used for assessing climate stability and global climate sensitivity (Fig. S5). During the first 100 years, the strong volcanic activity of the late 19th and early 20th century cools *historical* (blue curve) relative to *piControl*. Their difference reduces in the first half of the 20th century and anthropogenic warming emerges in the 70's when *historical* crosses *piControl*. At the start of the 21st century, *historical* is 0.7 K above its 1850–1900 mean and 0.5 K above *piControl*. While these values are lower than the 0.8 K observational estimate (e.g., Peters, 2016), the simulated warming towards the end of the 20th century matches the observed trend well and overestimates it during the beginning of the 21st century (blue versus black).

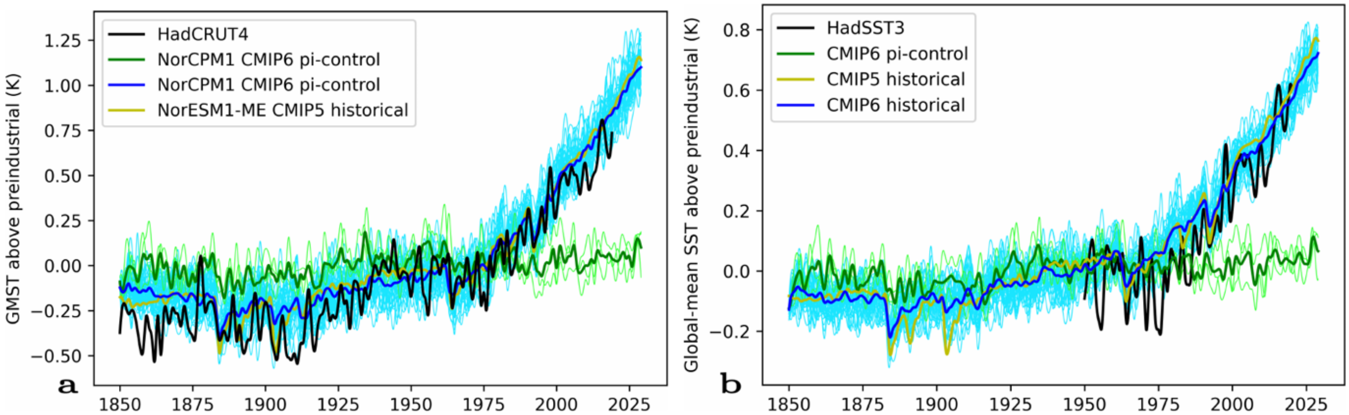
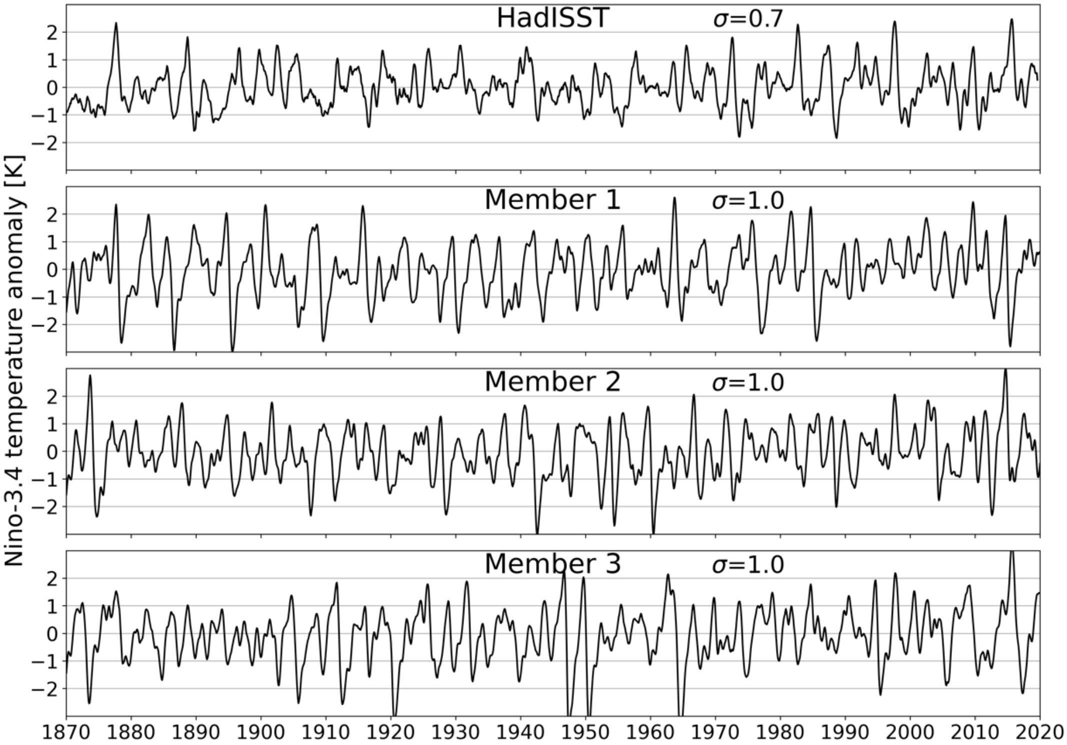


Figure S5: Global Mean SAT (left) and SST (right) evolutions. Thick lines denote ensemble means, thin lines individual members. CMIP6 data are shifted relative to the mean of CMIP6 *pi-control*. CMIP5 *historical* (30-member mean) is shifted to match the mean of CMIP6 *historical*. Observational estimate derived from HadCRUT4 (Morice et al., 2012) and HadSST 3.1.1 (Kennedy et al., 2011).

The forced GMST evolution of NorCPM1 behaves very similarly to the one of the CMIP5 NorESM1-ME system (blue versus yellow) despite the model and forcing upgrades. Deviations are most notable in the volcanic climate response at the start of the 20th century and are likely driven by changes in forcing rather than updates to the model. That updates to the model have a limited effect on the model's climate sensitivity is further confirmed with estimates of the effective climate sensitivity (ECS) to a doubling of atmospheric CO_2 relative to pre-industrial level. We estimate the ECS from the output of *abrupt4XCO2* using the linear regression method of Gregory et al. (2004). The 3.06 K estimate for NorCPM1 is close to the 2.94 K for NorESM1-ME. Both estimates derived from single simulations and are statistically indistinguishable given the large sampling uncertainty that results in 5–95% bootstrap confidence intervals of 2.99–3.19 K and 2.86–3.09 K for NorCPM1 and NorESM1-ME, respectively. These estimates are at the low end but well within the 2.1–4.7 K CMIP5 (Andrews et al., 2012) and 1.8–5.6 CMIP6 (Zelinka et al., 2020) model ranges and also within the narrowed observation-based range of 2.6–3.9 K (Sherwood et al., 2020). *piControl* (green curve) exhibits a small positive GMST long-term drift of 0.06 K century^{-1} that is caused by a 0.13 W m^{-1} net radiative transport at the top of atmosphere but is inconsequential for the sensitivity evaluation.

The model produces reasonably realistic ENSO variability in terms of frequencies and variance as diagnosed from the monthly Niño-3.4 index of *historical* (Fig. S6, S7). The simulated ENSO shows elevated power in the 2–5-year frequency band, similar to observations. The model overestimates the standard-deviation by 20 %.



100 **Figure S6: Niño-3.4 monthly index for HadISST (Rayner et al., 2003) and the first three members of NorCPM1's *historical* experiment.**

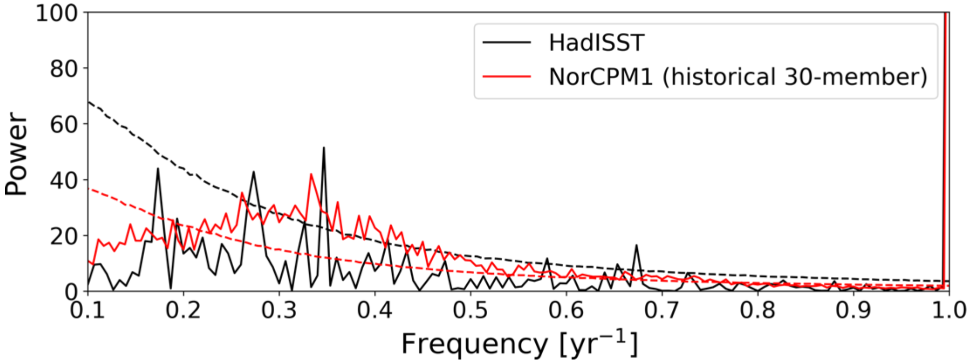


Figure S7: Niño-3.4 frequency spectrum. The observed estimate (black) uses monthly HadISST data (Rayner et al., 2003) over the period 1870-2019. The model estimate (red) uses monthly data from NorCPM1's *historical* experiment and represents the average of the spectra from the 30 members. The stippled lines show the 90th percentile from synthetic red noise time series with lag-1 105 **autocorrelation from the HadISST data. Differences in the two red noise estimates are due to the model data having 30 times as many data points compared to the observations.**

S2 Supporting reanalysis and hindcast evaluation

The different AMOC evolutions in *assim-i1* versus *assim-i2* indicate that NorCPM1's AMOC initialisation is sensitive to the climatology used for defining anomalies and/or to the assimilation update of sea ice. To further assess the relative importance of the two factors, we reran the *assim-i1* reanalysis, but this time using the climatological period 1950-2010 that was used in *assim-i2*. Before 1970, the AMOC evolution of the new reanalysis stays close to the strong AMOC evolution of *assim-i2* (Fig. S8). After 1970, however, it separates from *assim-i2* and stays somewhere between *assim-i1* and *assim-i2*. This suggests that the assimilation update of sea ice also plays a role in strengthening the AMOC in addition to the climatology, possibly by making the high-latitude oceans more saline (Fig. 3f).

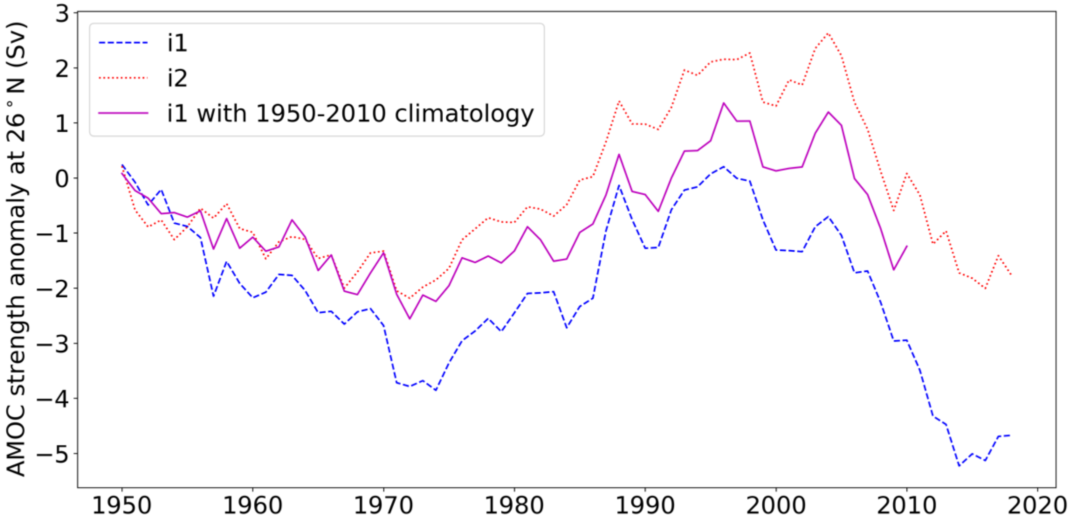


Figure S8: Attribution of AMOC differences to climatology versus sea ice assimilation. Annual-mean AMOC strength at 26 °N from the ensemble means of *assim-i1* (dotted red), *assim-i1b* (solid purple), and *assim-i2* (stippled blue). The *assim-i1b* reanalysis experiment is identical to *assim-i1* but uses the climatological period 1950–2010 (same as in *assim-i2*) for computing assimilation anomalies.

Skilful ENSO prediction is a likely driver of NorCPM1's robust initialisation benefit for first-year climate prediction. Figure S9 shows ACCs for the Nino-3.4 index over lead months. The ACCs of the initialised hindcasts stay above 0.8 from November until April, similar to *persistence* but considerably higher than the skill of *historical*. Following, the skill drops sharply, but in contrast to *persistence* the skill of the initialised hindcasts levels out and stays at 0.4-0.5 during summer and autumn, which is slightly above the skill of *historical*. During the subsequent winter and spring, the skill of the initialised hindcasts drops slightly, but much less than that of *historical*, suggesting a reemergence of initialisation benefit during the second spring. NorCPM1's results for November initialisation agree well with those Wang et al. (2019) obtained with NorCPM assimilating only SST observations (their Fig. 10). For initialization in other seasons, they show that NorCPM markedly outperforms *persistence*.

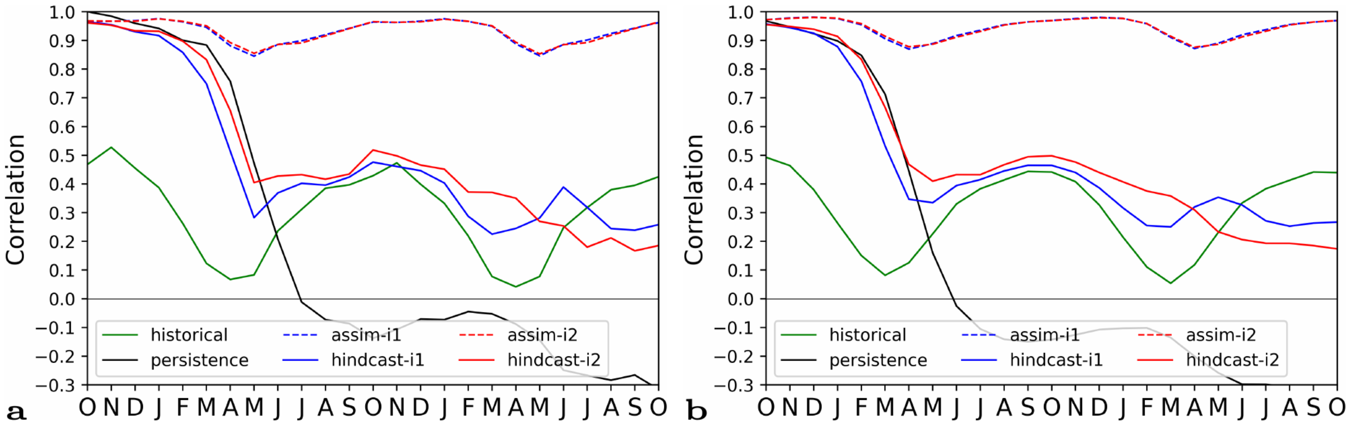


Figure S9: Prediction skill for Niño-3.4 index. (a) ACC over lead months with observations from HadISST (Rayner et al., 2003) for the period 1950–2019. The observed October value is used as persistence forecast. (b) As a, but applying a 3-month average prior to ACC computation and using the August–October average as persistence forecast.

While skilful SST prediction is key for skilful prediction in atmosphere and over land, the consideration of ocean heat content can provide additional clues on its relation to ocean thermal inertia and ocean dynamics. Because the ocean subsurface state is less directly influenced by unpredictable atmospheric or coupled variability, one would expect to see higher skill there.

The ACCs for 0-300 m temperature (Fig. S10) have a similar pattern as those of SST (Fig. 12) but tend to be smaller and patchier. A caveat is that our observational reference for subsurface temperature is less reliable than that for SST over the 1960-2018 evaluation period. The skill increment from initialisation in SST appears related to the upper ocean heat content (Fig 12d,i,n; Fig D3d,i,n).

Contrary to temperature, advected salinity signals are not subject to atmospheric feedback-driven surface damping. Predictions of upper ocean salinity can therefore inform temperature predictions by providing additional insights. The ACCs for 0-300 m salinity (Fig. S11) have generally a similar pattern but are lower than those of temperature. Lower coverage of salinity compared to temperature profiles and a weak DA constraint of mixed layer salinity may partly explain the lower ACCs. Exploiting the profile information of the mixed layer would potentially improve the skill. The absolute ACCs suggest skilful multiyear prediction of salinity over large regions (Fig. S11a,f,k). However, there are fewer places where the initialized dynamical predictions outperform both the persistence forecast (Fig. S11c,h,m) and uninitialized prediction (Fig. S11d,i,n).

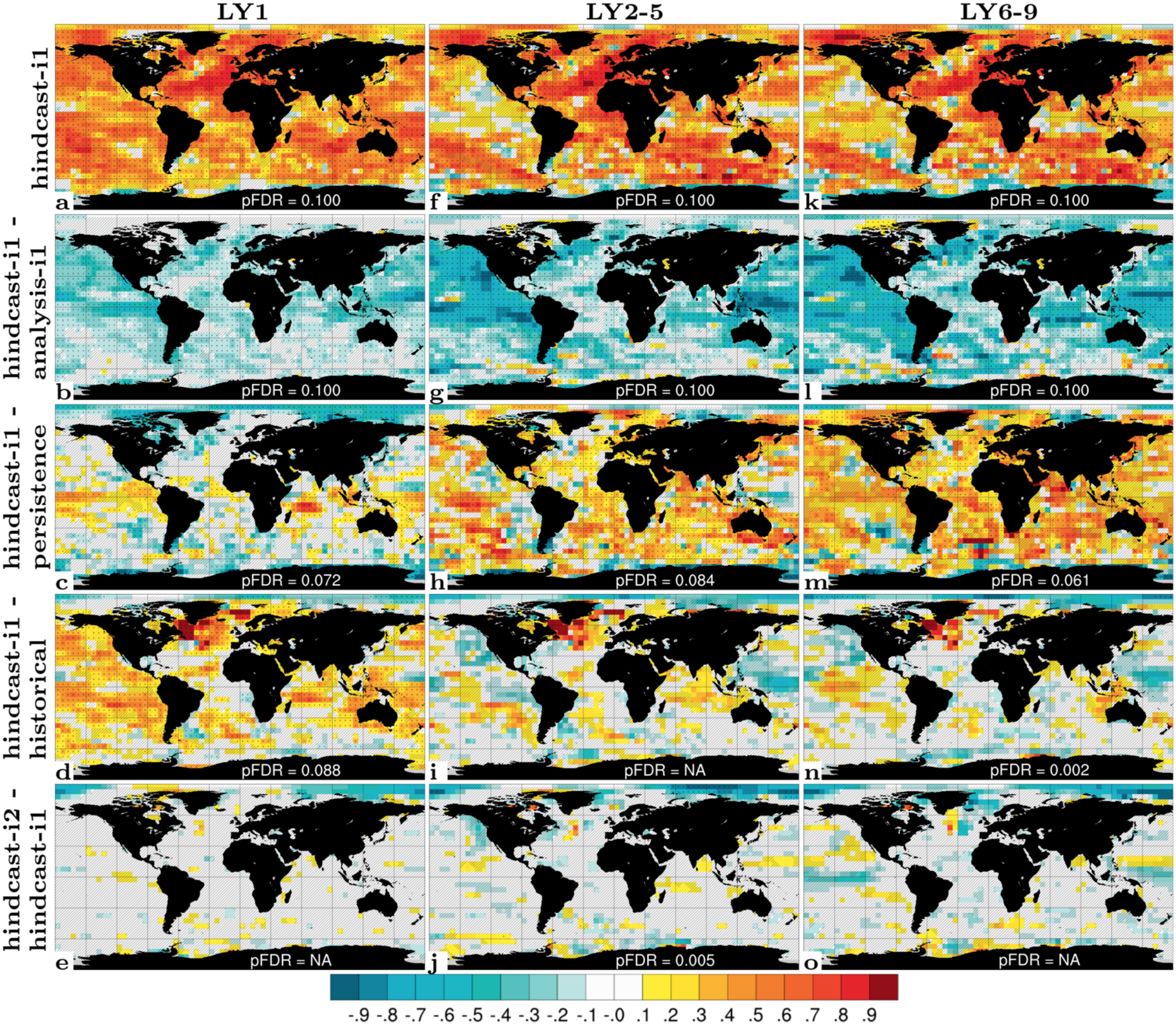


Figure S10: Prediction skill for 0-300 m temperature (T300). ACC of *hindcast-i1* (a), ΔACC of *hindcast-i1 - analysis-i1* (b), *hindcast-i1 - persistence* (c), *hindcast-i1 - historical* (d), and *hindcast-i2 - hindcast-i1* (e) for LY1. Middle and right column show the same but for LY2–5 (f–j) and LY6–9 (k–o). Observations use EN4.2.1 (Good et al., 2013) with coverage 1960–2018. Hatched areas are not locally significant, dotted areas are field significant. 150

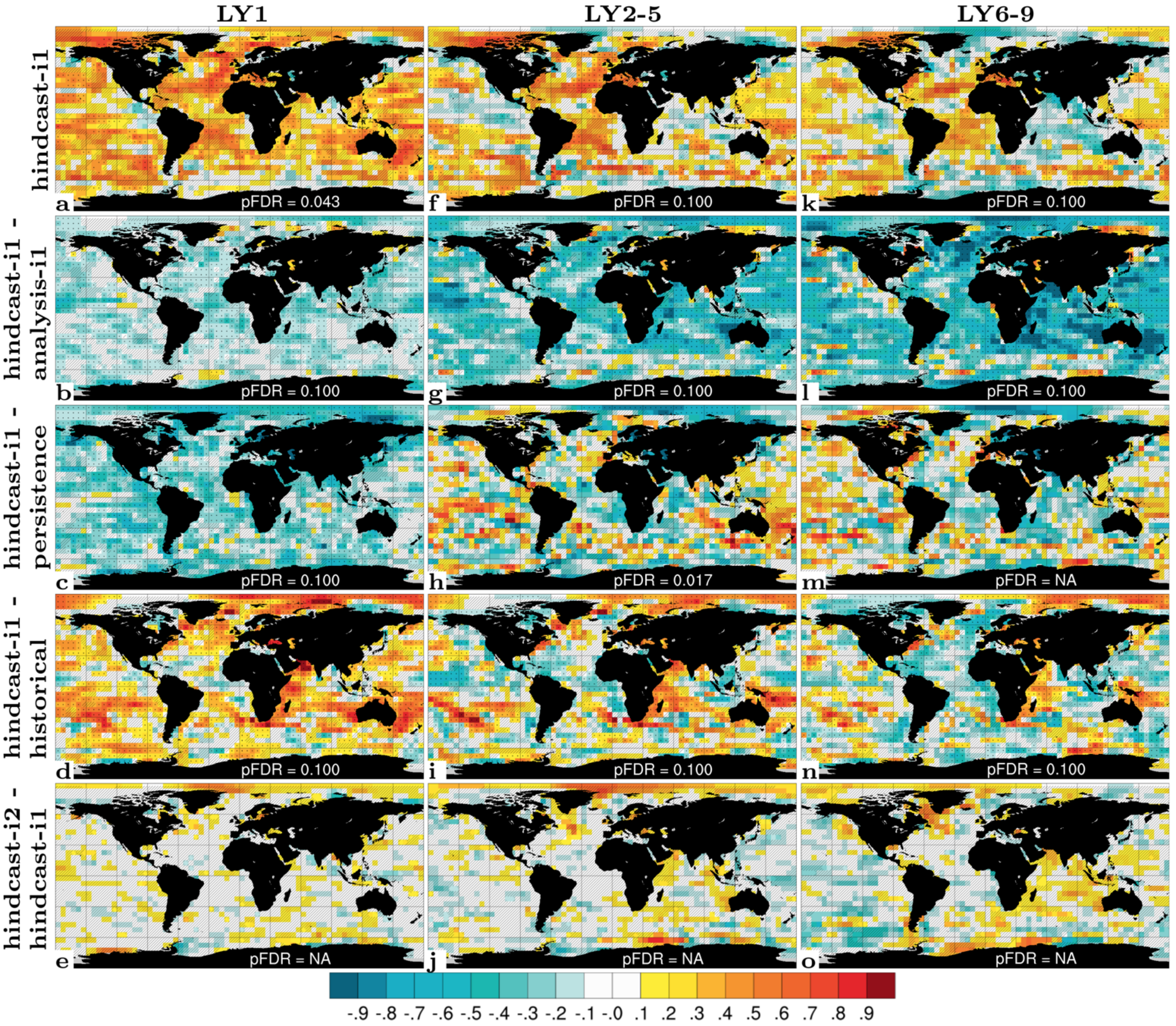


Figure S11: Prediction skill for 0-300 m salinity (S300). ACC of *hindcast-i1* (a), ΔACC of *hindcast-i1 - analysis-i1* (b), *hindcast-i1 - persistence* (c), *hindcast-i1 - historical* (d), and *hindcast-i2 - hindcast-i1* (e) for LY1. Middle and right column show the same but for LY2–5 (f–j) and LY6–9 (k–o). Observations use EN4.2.1 (Good et al., 2013) with coverage 1960–2018. Hatched areas are not locally significant, dotted areas are field significant. 155

Skilful decadal predictions of sea level changes has societal value, but the evaluation is complicated by the short satellite record. Nevertheless, the ACC maps for SSH (Fig. S12) reveal that most of the skill comes from the externally forced trend.

The results indicate some benefit from initialisation in the SPNA, but degradation at the inter-gyre region and from there extending into the Nordic Seas. *hindcast-i2* performs better than *hindcast-i1* in the high-latitudes.

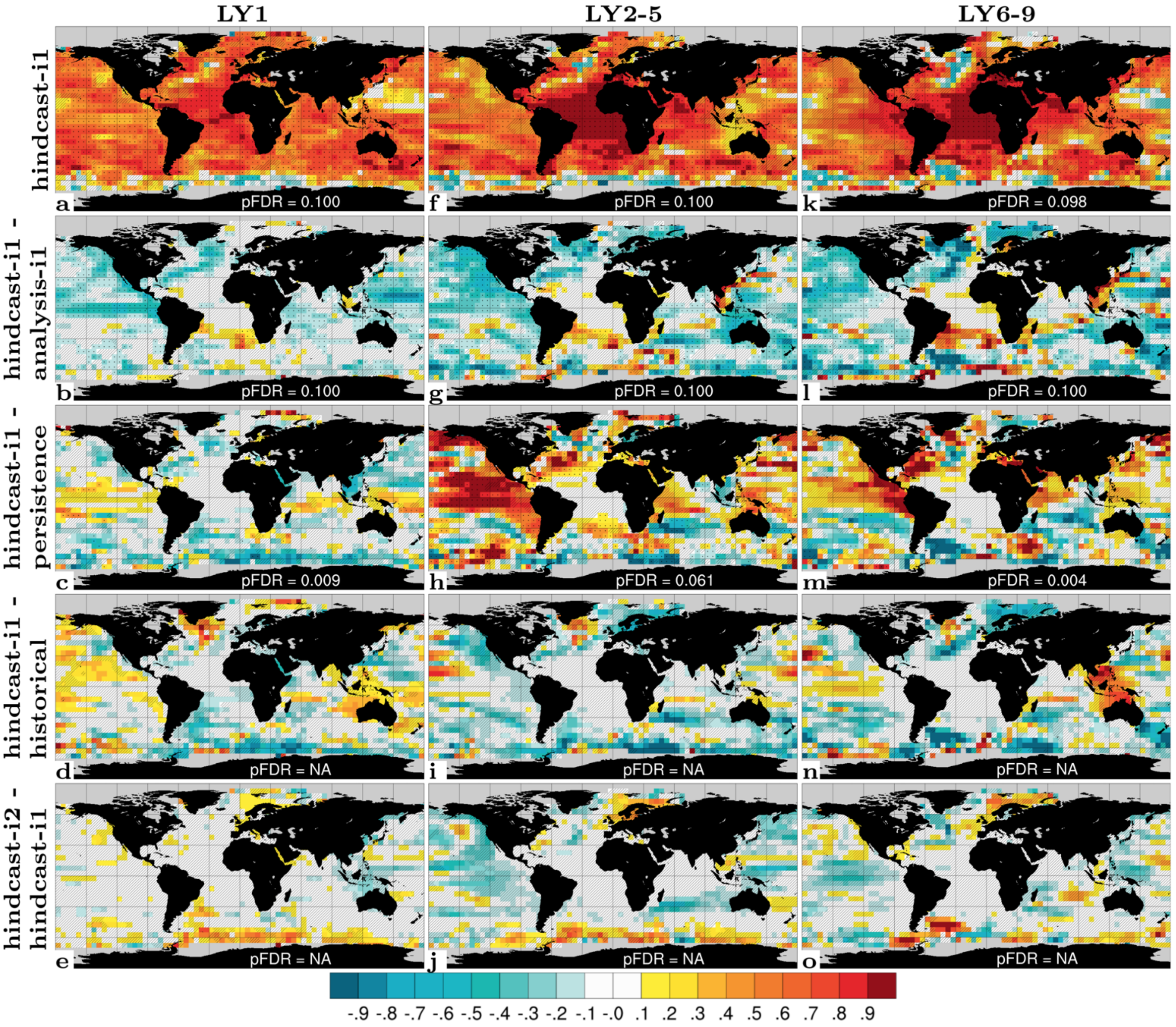


Figure S12: Prediction skill for sea surface height (SSH). ACC of *hindcast-i1* (a), ΔACC of *hindcast-i1 - analysis-i1* (b), *hindcast-i1 - persistence* (c), *hindcast-i1 - historical* (d), and *hindcast-i2 - hindcast-i1* (e) for LY1. Middle and right column show the same but for LY2–5 (f–j) and LY6–9 (k–o). Observations use ARMOR-3D (Larnicol et al., 2006) with coverage 1993–2018. Hatched areas are not locally significant, dotted areas are field significant.

SPNA temperature variability has been linked to meridional ocean heat transport and overturning variability. This relation holds also for the SPNA temperature evolutions in NorCPM1's hindcast (Fig. 13a,d). These can be well approximated from anomalous northward ocean heat transport into the SPNA region (Fig. S13a,c) and from regressing this transport on AMOC strength (Fig. S13b,d). That the diagnosed temperature changes overestimate the amplitude of the simulated temperature changes is expected, because the surface heat flux acts to dampen the SPNA temperature anomalies in the model.

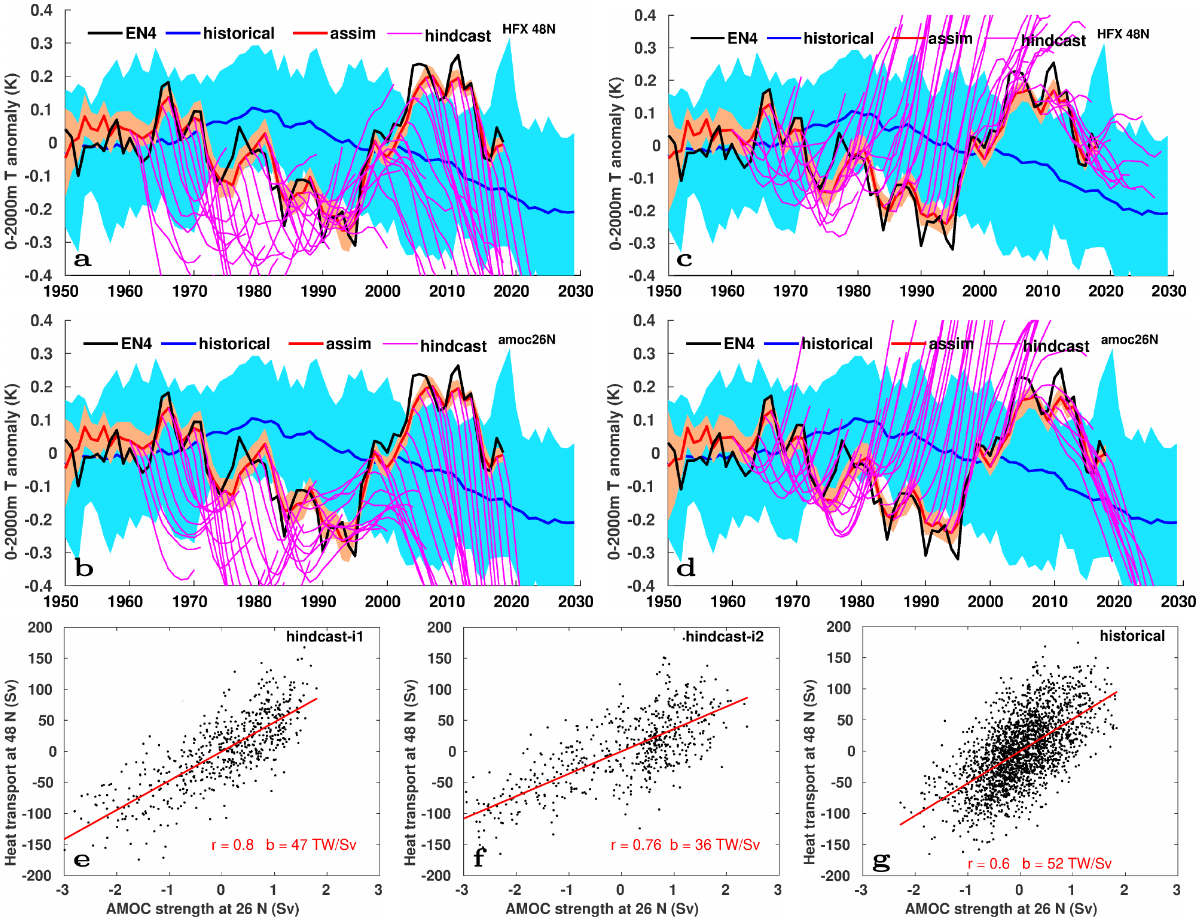


Figure S13: SPNA [60-15 °W, 48-65 °N] 0–2000 m temperature (T2000) evolution estimated from anomalous heat transport across 48 °N (HFX48) (a,c) and from anomalous AMOC strength at 26 °N (MOC26) (b,d) for i1 and i2. Solid lines show ensemble means of *historical* (blue), *assim* (red), and *hindcast* (purple) experiments, with the 1950–2010 average of historical subtracted. Shading denotes ensemble minima and maxima. Estimates shown for *hindcast*, actual temperatures for other experiments. Bottom panels show HFX48 over MOC26 with correlation and regression coefficient from annual values of hindcast-i1 (e) and hindcast-i2 (f) and 5-year averages of historical (g). A factor of 50 TW/Sv has been applied for deriving HFX48 from MOC26 in b,d. The conversion from HFX48 to T2000 tendency disregards temperature changes below 2000 m and heat transport changes through the western-eastern-northern boundaries as well as surface of the SPNA domain.

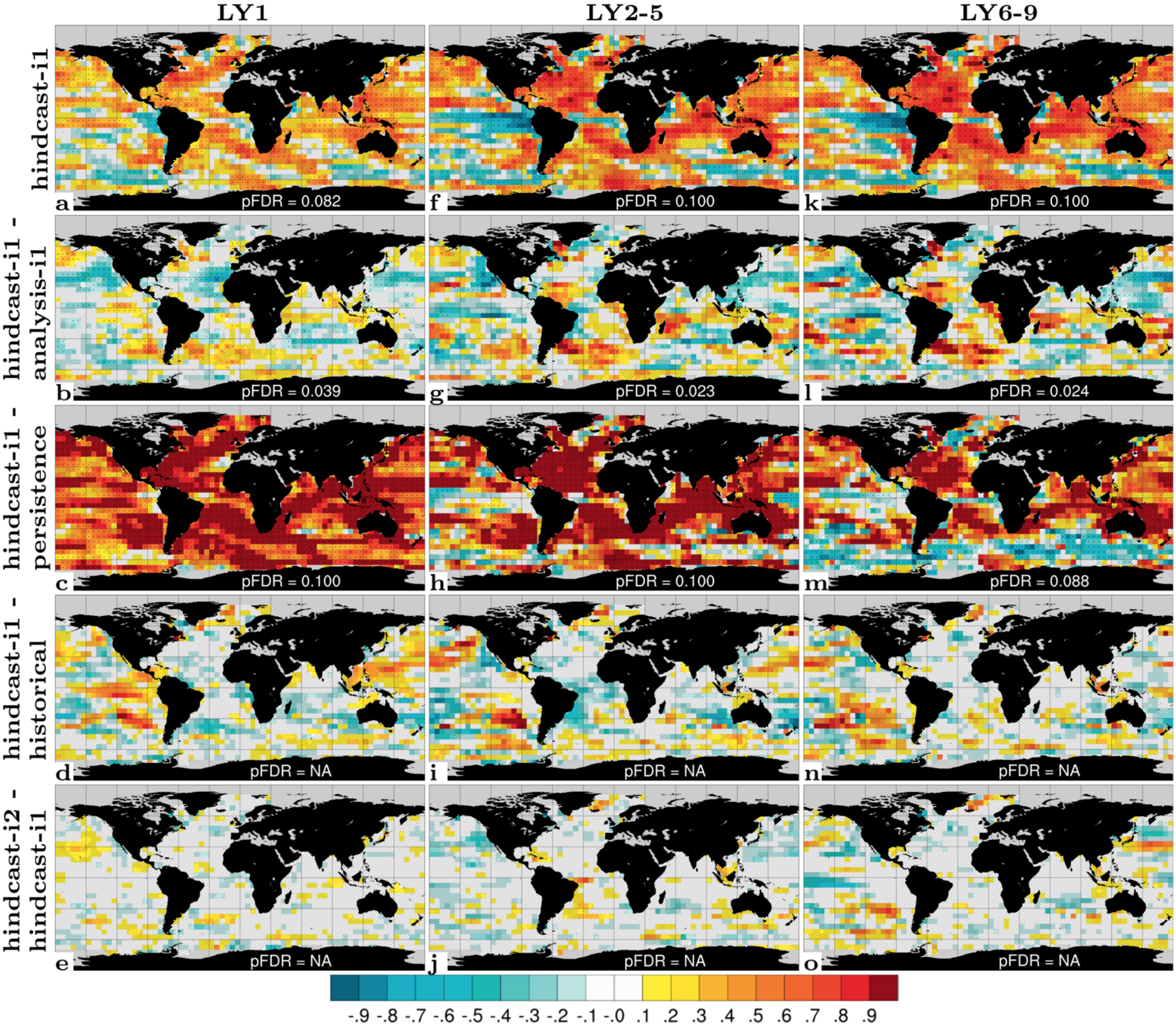


Figure S14: Prediction skill for surface CO_2 flux. ACC of *hindcast-i1* (a), ΔACC of *hindcast-i1 - analysis-i1* (b), ΔACC of *hindcast-i1 - persistence* (c), ΔACC of *hindcast-i1 - historical* (d) and ΔACC of *hindcast-i2 - hindcast-i1* (e) for LY1. Middle and right column show the same but for LY2–5 (f–j) and LY6–9 (k–o). Observations use SOCCOM (Landschützer et al., 2019) with coverage 1982–2017. Hatched areas are not locally significant, dotted areas are field significant.

For the linearly detrended surface CO_2 fluxes, a high total skill is found for all lead year ranges in the Atlantic, Indian and in the North Pacific Oceans (Fig. S14). Most of the skill comes from the external forcings, as revealed by small ΔACCs (of both signs) for *hindcast-i1* - *historical* (Fig. S14d,i,n). The relatively high ΔACCs in the southern subtropical Pacific are a result of subtracting ACCs that are negative in *historical*. Only in the tropical Pacific and for LY1 initialisation the initialization benefit translates into moderately positive total skill.

Anomaly initialisation has the advantage that it minimizes forecast drift. The decadal evolutions of the integrated Arctic sea ice volume (Fig. S15) illustrate that the sea ice initialisation for *hindcast-i2* changes the mean state and therefore has a full-field character, causing considerable forecast drift, despite the use of anomaly assimilation. Posteriori drift correction can thus be necessary also for prediction systems that use anomaly assimilation.

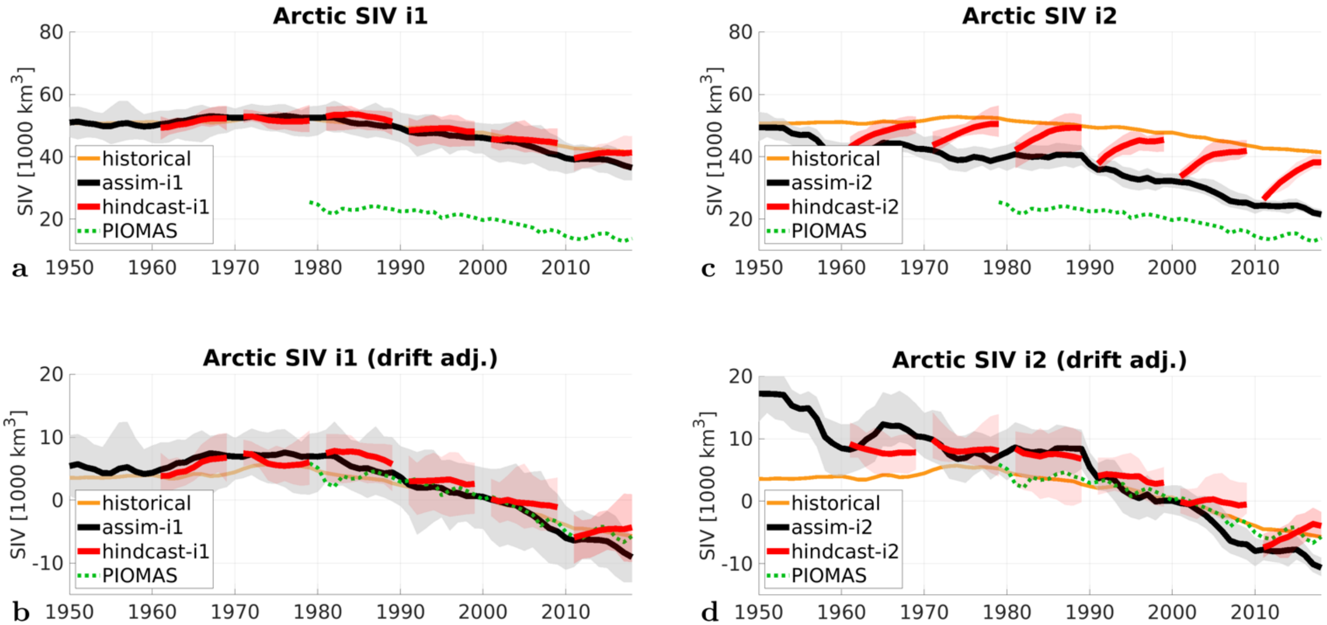


Figure S15: Time-development of annual Arctic sea ice volume (SIV) of experiment i1 (left) and i2 (right). Black lines are ensemble averages of the respective reanalyses, grey shadings depict the ensemble envelope. Red lines show ensemble averages of the respective decadal hindcasts, red shadings the ensemble envelope, for hindcasts branched in the beginning of each decade. Annual SIV derived from PIOMAS (green) and annual SIV of the ensemble mean of the historical run (orange) are provided as reference. The bottom row shows the drift-adjusted SIV determined as in Yeager et al. (2018) using 1980–2018 as reference period for the adjustment.

We assess the skill in predicting atmospheric circulation variability from Z500 (Fig. S16) and SLP (Fig. S17). For Z500, *hindcast-i1* exhibits positive LY1 skill over most regions (Fig. S16a) that is considerably higher than persistence skill (Fig. S16c). Initialisation benefit related to ENSO is found over the tropical Pacific and also over the Southern Ocean, extratropical North Pacific and subpolar North Atlantic, and over the Canadian Arctic (Fig. S16d). For LY2–5 and LY6–9, the absolute skill (Fig. S16f,k) is higher than for LY1 and close to saturation in some regions. Most of the LY1 initialisation benefit has disappeared, but some remains over the subpolar North Atlantic and central Siberia (Fig. S16i,n) where the absolute skill of *hindcast-i1* is low and variability is dominated by the North Atlantic Oscillation. The assimilation update of sea ice in *hindcast-i2* slightly benefits the multiyear skill over the Arctic (Fig. S16j,o), offsetting some of the skill degradation caused by initialisation in *hindcast-i1* (Fig. S16i,n).

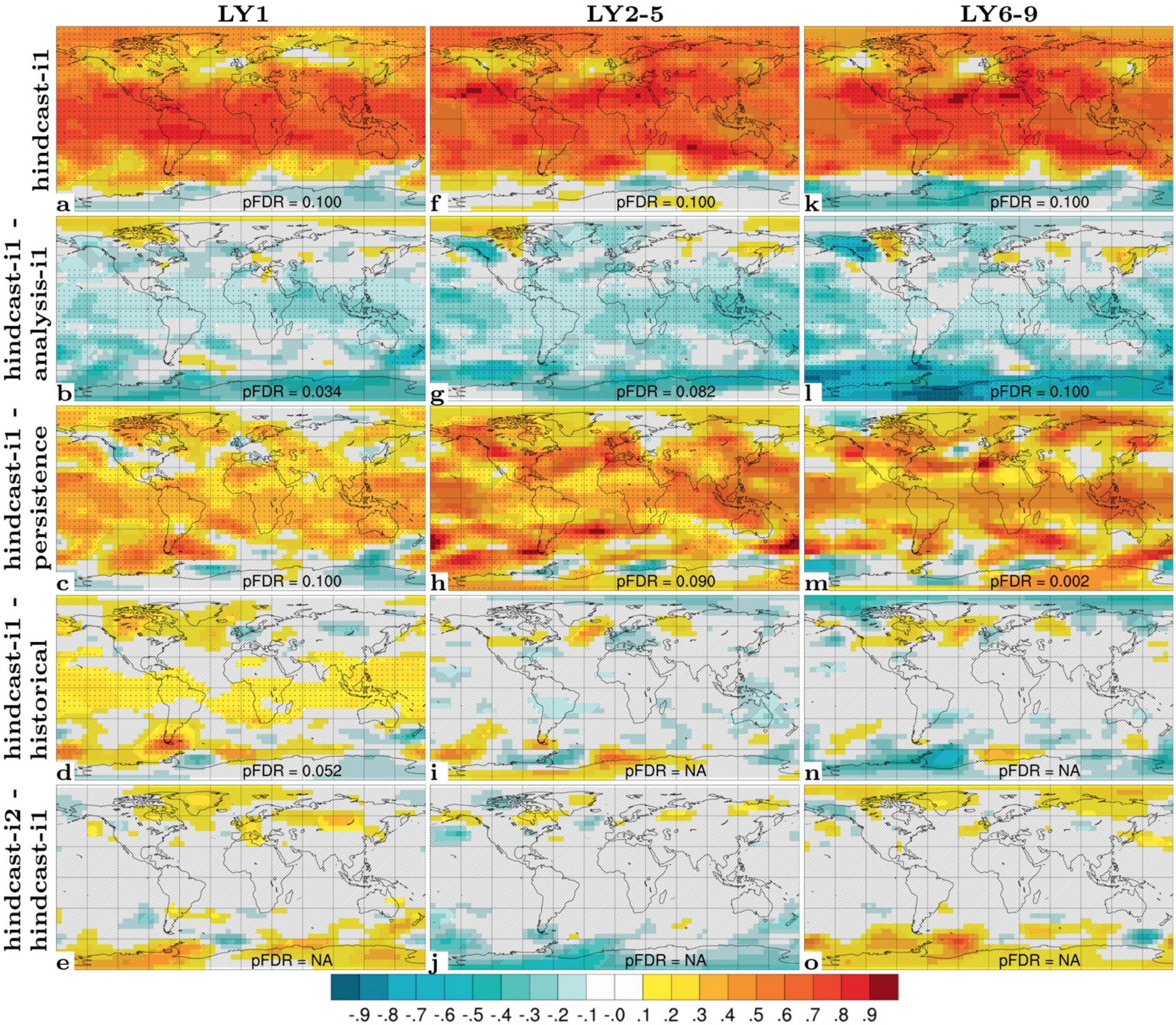


Figure S16: Prediction skill of 500 hPa geopotential height (Z500). ACC of *hindcast-i1* (a), ΔACC of *hindcast-i1 - analysis-i1* (b), *hindcast-i1 - persistence* (c), *hindcast-i1 - historical* (d), and *hindcast-i2 - hindcast-i1* (e) for LY1. Middle and right column show the same but for LY2–5 (f–j) and LY6–9 (k–o). Observations use extended ERA5 reanalysis (Hersbach et al., 2020) with coverage 1950–2019. Hatched areas are not locally significant, dotted areas are field significant.

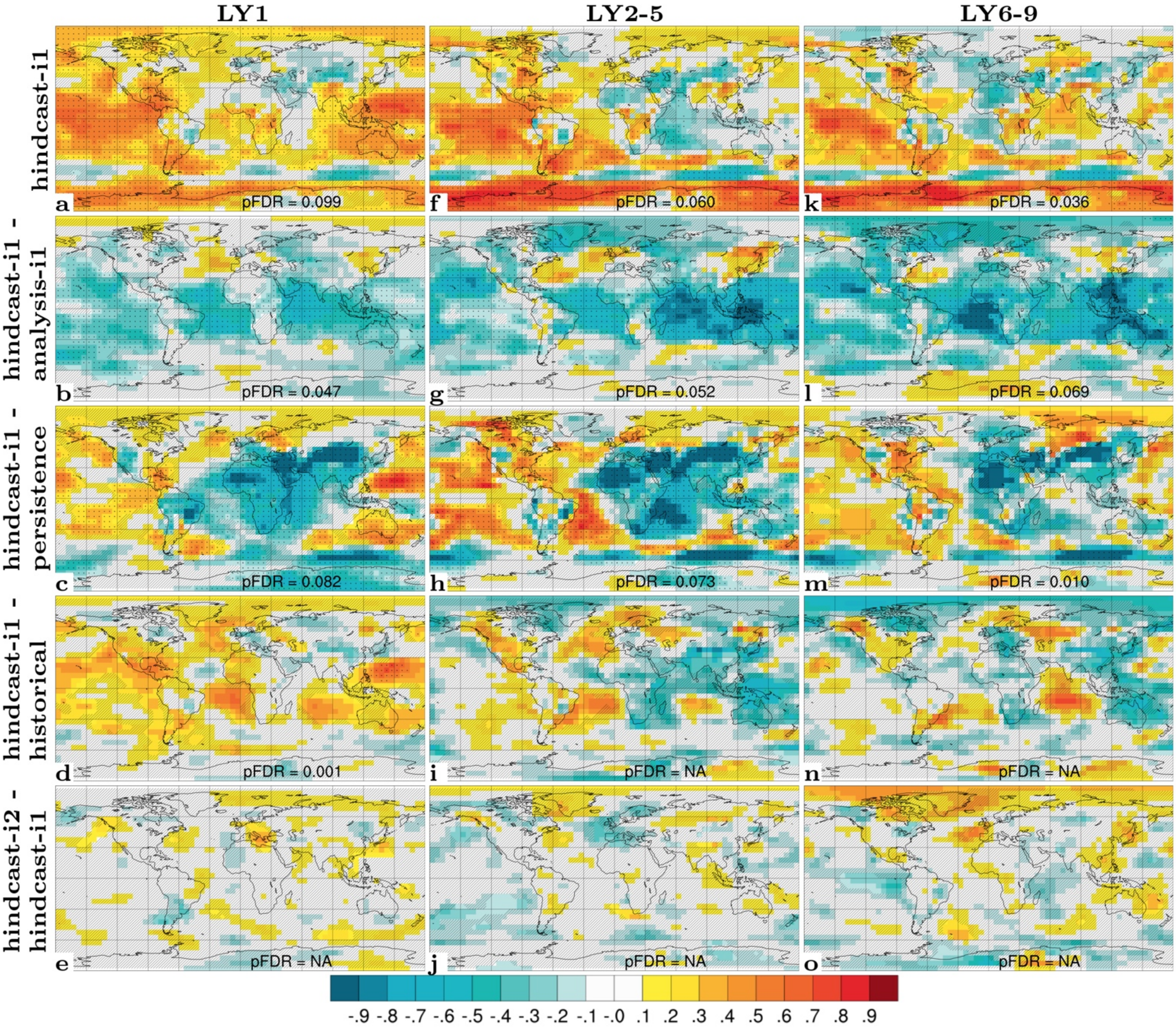


Figure S17: Prediction skill of sea level pressure (SLP). ACC of *hindcast-i1* (a), ΔACC of *hindcast-i1 - analysis-i1* (b), *hindcast-i1 - persistence* (c), *hindcast-i1 - historical* (d), and *hindcast-i2 - hindcast-i1* (e) for LY1. Middle and right column show the same but for LY2–5 (f–j) and LY6–9 (k-o). Observations use NCEP reanalysis (Kalnay et al., 1996) with coverage 1950–2019. Hatched areas are not locally significant, dotted areas are field significant.

For SLP, *hindcast-i1* exhibits positive LY1 skill over most regions except Central Eurasia, Brazil and the Southern Ocean (Fig. S17a) and generally performs better than uninitialized prediction (Fig. S17d). The benefit from initialisation is less evident for LY2–5 and LY6–9 (Fig. S17i,n), showing roughly equal proportions of positive and negative ΔACCs. For all lead year ranges, *hindcast-i1* performs better than persistence in the Pacific and western Atlantic sectors and considerably worse in the African/Indian sector, something that warrants further investigation. The assimilation update of sea ice in *hindcast-i2* again benefits the multiyear skill over the Arctic (Fig. S17j,o), offsetting some of the skill degradation caused by initialisation in *hindcast-i1* (Fig. S17i,n).

The use of ACC differences for detecting benefit from initialization can be problematic if the skill from the externally forced climate trend is very high and the ACC differences are small (but can nevertheless reflect a significant change in explained variance). Smith et al. (2019) proposed a more robust way of evaluating initialization benefit by regressing out the forced signal of the model—estimated from the ensemble mean of a historical simulation experiment with the same model—from both model and observation output and then computing the ACC of the residuals. The result is scaled with the standard-deviation of the residuals of the observations divided by the standard-deviation of the observations, and an estimated spurious correlation bias is subtracted. Figure S18 compares the ΔACC versus the residual ACC results obtained for SAT. Initialization benefit over land not shown by the ΔACCs is robustly detected by the residual ACCs.

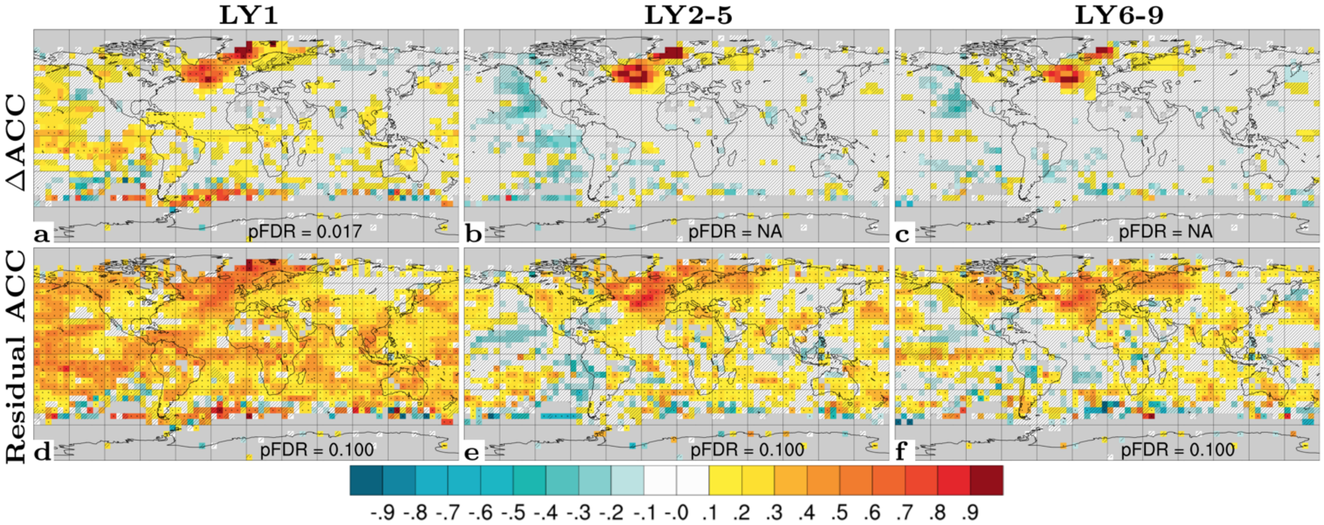


Figure S18: Initialisation benefit for 2 m temperature (SAT) computed as ΔACC versus ACC of residuals. ΔACC of *hindcast-i1* - *historical* for LY1 (a), LY2–5 (b), and LY6-9 (c). ACC of residuals for LY1 (d), LY2–5 (e) and LY6-9 (f). The residuals are computed by regressing out the ensemble mean of *historical*—i.e., the forced signal of the model—from hindcast-i1 and the observations. Following Smith et al. (2019), the ACCs are scaled with the standard-deviation of the observation residuals divided by standard-deviation of the observations, and an estimated spurious correlation bias is subtracted. Observations use HadCRUT4 (Morice et al., 2012) with coverage 1950–2019. Hatched areas are not locally significant, dotted areas are field significant.

The Mean Square Skill Score (MSSS; Goddard et al., 2013) is an alternative metric to the ACC and computed as 1 - MSE/MSE$_{\mathrm{ref}}$, where MSE$_{\mathrm{ref}}$ is the Mean Square Error (MSE) of a reference prediction (e.g., climatology, persistence forecast, historical simulation, or other). Unlike the ACC, the MSSS penalizes amplitude errors and has a range from -∞ to 1, with positive values indicating that the predictions outperform the reference prediction. The MSSS results for T300 (Fig. S19) are qualitatively similar to the ACC results (Fig. S3).

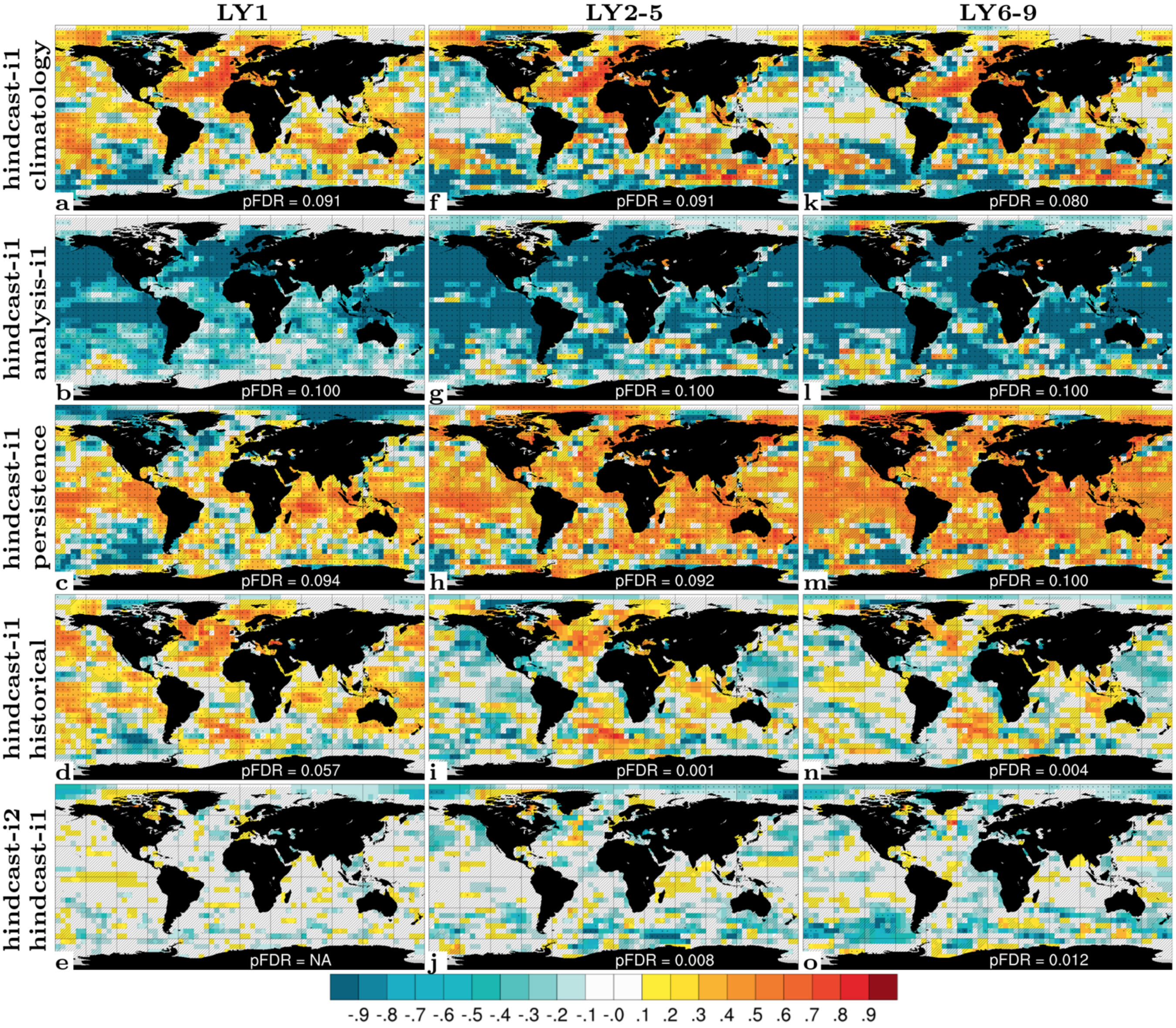


Figure S19: Prediction skill for 0-300 m temperature (T300). MSSS of hindcast-i1 (a), MSSS of *hindcast-i1* with respect to *analysis-i1* (b), *hindcast-i1* with respect to *persistence* (c), *hindcast-i1* with respect to *historical* (d), and *hindcast-i2* with respect to *hindcast-i1* (e) for LY1. Middle and right column show the same but for LY2–5 (f–j) and LY6–9 (k–o). Observations use EN4.2.1 (Good et al., 2013) with coverage 1960–2018. Hatched areas are not locally significant, dotted areas are field significant. The MSSS calculation follows Yeager et al. (2018).

Supporting references

Andrews, T., Gregory, J. M., Webb, M. J., and Taylor, K. E.: Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models, Geophysical Research Letters, 39, 2012.

Bentsen, M., Bethke, I., Debernard, J. B., Iversen, T., Kirkevåg, A., Seland, Ø., Drange, H., Roelandt, C., Seierstad, I. A., Hoose, C., and Kristjánsson, J. E.: The Norwegian Earth System Model, NorESM1-M – Part 1: Description and basic evaluation of the physical climate, Geosci. Model Dev., 6, 687-720, https://doi.org/10.5194/gmd-6-687-2013, 2013.

Fasullo, J. T. and Trenberth, K. E.: The Annual Cycle of the Energy Budget. Part I: Global Mean and Land–Ocean Exchanges, Journal of Climate, 21, 2297–2312, https://doi.org/10.1175/2007JCLI1935.1, 2008.

Goddard, L., Kumar, A., Solomon, A., Smith, D., Boer, G., Gonzalez, P., Kharin, V., Merryfield, W., Deser, C., Mason, S. J., et al.: A verification framework for interannual-to-decadal predictions experiments, Climate Dynamics, 40, 245–272, https://doi.org/10.1007/s00382-012-1481-2, 2013.

Good, S., Martin, M. J., and Rayner, N.: EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates, Journal of Geophysical Research: Oceans, 118, 6704–6716, https://doi.org/10.1002/2013JC009067, 2013.

Gregory, J. M., Ingram, W., Palmer, M., Jones, G., Stott, P., Thorpe, R., Lowe, J., Johns, T., and Williams, K.: A new method for diagnosing radiative forcing and climate sensitivity, Geophysical research letters, 31, https://doi.org/10.1029/2003GL018747, 2004.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne, M. J., Smith, T. M., Vose, R. S., and Zhang, H.-M.: Extended reconstructed sea surface temperature, version 5 (ERSSTv5): upgrades, validations, and intercomparisons, Journal of Climate, 30, 8179–8205, https://doi.org/10.1175/JCLI-D-16-0836.1, 2017.

Iversen, T., Bentsen, M., Bethke, I., Debernard, J. B., Kirkevåg, A., Seland, Ø., Drange, H., Kristjansson, J. E., Medhaug, I., Sand, M., and Seierstad, I. A.: The Norwegian Earth System Model, NorESM1-M – Part 2: Climate response and scenario projections, Geoscientific Model Development, 6, 389–415, https://doi.org/10.5194/gmd-6-389-2013, 2013.

Kennedy, J. J., Rayner, N. A., Smith, R. O., Parker, D. E., and Saunby, M.: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization, Journal of Geophysical Research: Atmospheres, 116, https://doi.org/10.1029/2010JD015220, 2011.

Landschützer, P., Bushinsky, S., and Gray, A. R.: A combined globally mapped CO2 flux estimate based on the Surface Ocean CO2 Atlas Database (SOCAT) and Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) biogeochemistry floats from 1982 to 2017 (NCEI Accession 0191304). Version 1.1, NOAA National Centers for Environmental Information. Dataset., https://doi.org/https://doi.org/10.25921/9hsn-xq82, 2019.

Larnicol, G., Guinehut, S., Rio, M. H., Drévillon, M., Faugere, Y., and Nicolas, G.: The Global Observed Ocean Products of the French Mercator Project, ESA Special Publication, 614, 110, 2006.

Morice, C. P., Kennedy, J. J., Rayner, N. A., and Jones, P. D.: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, Journal of Geophysical Research: Atmospheres, 117, https://doi.org/10.1029/2011JD017187, 2012.

Passos, L. G., Langehaug, H. R., Årthun, M., Eldevik, T., Bethke, I., and Kimmritz, M.: Impact of initialization techniques on the predictive skill of the Arctic-Atlantic region in the Norwegian Climate Prediction Model (NorCPM), Climate Dynamics, https://doi.org/10.21203/rs.3.rs-766415/v1, under review.

Peters, G. P.: The 'best available science' to inform 1.5 °C policy choices, Nature Climate Change, 6, 646, https://doi.org/10.1038/nclimate3000, 2016.

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

Reynolds, R. W., Rayner, N. A., Smith, T. M., Stokes, D. C., and Wang, W.: An Improved In Situ and Satellite SST Analysis for Climate, Journal of Climate, 15, 1609–1625, https://doi.org/10.1175/1520-0442(2002)015<1609:AIISAS>2.0.CO;2, 2002.

Sherwood, S., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., Hegerl, G., Klein, S. A., Marvel, K. D., Rohling, E. J., et al: An assessment of Earth's climate sensitivity using multiple lines of evidence, Reviews of Geophysics, 58, https://doi.org/10.1029/2019RG000678, 2020.

Smith, D., Eade, R., Scaife, A. A., Caron, L.-P., Danabasoglu, G., DelSole, T., Delworth, T., Doblas-Reyes, F., Dunstone, N., Hermanson, L., et al.: Robust skill of decadal climate predictions, Npj Climate and Atmospheric Science, 2, 1–10, https://doi.org/10.1038/s41612-019-0071-y, 2019.

Yeager, S., Danabasoglu, G., Rosenbloom, N., Strand, W., Bates, S., Meehl, G., Karspeck, A., Lindsay, K., Long, M., Teng, H., et al.: Predicting near-term changes in the Earth System: A large ensemble of initialized decadal prediction simulations using the Community Earth System Model, Bulletin of the American Meteorological Society, 99, 1867–1886, https://doi.org/10.1175/BAMS-D-17-0098.1, 2018.

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., Klein, S. A., and Taylor, K. E.: Causes of higher climate sensitivity in CMIP6 models, Geophysical Research Letters, 47, e2019GL085 782, https://doi.org/10.1029/2019GL085782, 2020.