

<https://doi.org/10.1038/s41612-025-01090-0>

ACE2: accurately learning subseasonal to decadal atmospheric variability and forced responses



Oliver Watt-Meyer¹ ✉, Brian Henn¹, Jeremy McGibbon¹, Spencer K. Clark^{1,2}, Anna Kwa¹, W. Andre Perkins¹, Elynn Wu¹, Lucas Harris² & Christopher S. Bretherton¹

Existing machine learning models of weather variability are not formulated to enable assessment of their response to varying external boundary conditions such as sea surface temperature and greenhouse gases. Here we present ACE2 (Ai2 Climate Emulator version 2) and its application to reproducing atmospheric variability over the past 80 years on timescales from days to decades. ACE2 is a 450M-parameter autoregressive machine learning emulator, operating with 6-hour temporal resolution, 1° horizontal resolution and eight atmospheric vertical layers. It exactly conserves global dry air mass and moisture and can be stepped forward stably for arbitrarily many steps with a throughput of about 1500 simulated years per wall clock day. ACE2 generates emergent phenomena such as tropical cyclones, the Madden Julian Oscillation, and sudden stratospheric warmings. Furthermore, it accurately reproduces the atmospheric response to El Niño variability and global trends of temperature over the past 80 years. However, its sensitivities to separately changing sea surface temperature and carbon dioxide are not entirely realistic.

Machine learning offers an avenue to accelerate existing climate models by orders of magnitude. This acceleration is achieved by running efficiently on GPU hardware and by taking relatively long time steps, enabled by the lack of stability constraints that accompany traditional numerical methods. This increased efficiency has the potential to dramatically accelerate research tasks requiring many years of simulation. For example, it would enable easier exploration of large ensembles and rare events^{1,2} and allow accurate separation of forced response versus internal variability³. It would permit the lengthy simulations necessary for the study of paleoclimate with more realistic models than intermediate complexity models⁴. Finally, it would enable easy interpolation between wide range of climate change scenarios⁵. The cheap cost of inference and ability to run on consumer hardware opens the door of running climate models to a wider range of users. In addition to acceleration, a machine-learning based climate model emulator is differentiable, making it potentially useful for data assimilation applications^{6–8}.

The extent to which machine learning will lead to more accurate climate models remains to be seen. While machine learning has demonstrated an ability to improve weather prediction accuracy^{9–13}, the typical goal of climate prediction is to forecast previously unseen conditions, for example the expected global warming from a doubling of CO₂ concentration. Out-of-sample generalization is a fundamental challenge for machine learning, potentially necessitating the use of physics-based priors^{14,15} and the

training of machine learning based climate emulators on output from physics-based numerical models¹⁶. In this study we focus on emulating the climate of the historical period 1940–2020, including variability and trends. We demonstrate that our emulator can be skillfully trained on the ERA5 reanalysis¹⁷ or on an AMIP-style¹⁸ historical simulation with GFDL's SHiELD model¹⁹. SHiELD can also simulate perturbed climates, as would be needed to train an emulator that could be expected to simulate long-term climate change²⁰.

For this work, we use the Ai2 Climate Emulator version 2 (ACE2; see <https://github.com/ai2cm/ace>), a significant update to the ACE atmospheric model emulator described in refs. 21,22. Briefly, the emulator operates at 1° horizontal resolution with eight terrain-following vertical layers. It is initialized from a snapshot of atmospheric temperature, humidity and winds and can stably integrate forward an arbitrary number of 6-h time steps with a user-specified sea surface temperature (SST) boundary condition. The main methodological advances of ACE2 over version 1 of ACE are: (1) addition of CO₂ as a forcing variable, (2) ability to emulate observed atmospheric trends of the preceding 80 years and (3) the exact conservation of dry air mass and atmospheric moisture in ACE2 simulations. In addition, ACE2 is trained on two datasets to demonstrate its general applicability: first on an AMIP-style¹⁸ simulation with GFDL's SHiELD model¹⁹ and second on the ERA5 reanalysis¹⁷.

¹Allen Institute for Artificial Intelligence (Ai2), Seattle, WA, USA. ²Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, NJ, USA.

✉ e-mail: oliverwm@allenai.org

This study provides a more multifaceted evaluation of ACE2 than in our prior work on ACE, which only used annually-repeating climatological SSTs and sea-ice concentration. We show ACE2's accurate atmospheric response to El Niño variability as well as the long-term trends and inter-annual variability of global mean temperature and total water path. The ERA5-trained model allows evaluation of weather forecast skill and of phenomena such as tropical cyclones and the Madden Julian Oscillation, which are less well represented in the relatively coarse atmospheric models previously used for training ACE.

Related work includes NeuralGCM¹³ which showed some 30-year simulations with reasonable trends and low climate biases. However about one third of NeuralGCM's simulations went unstable before reaching 30 years, which limits its current applicability to climate prediction. Atmospheric emulators with long-term stability trained on ERA5^{23–25} and atmospheric model output^{21,22,26} have been reported, but none to date demonstrate the ability to accurately respond to the changing external forcing of the atmosphere over the last 80 years.

Results

Training period evaluation

We present ACE2 model evaluations initialized in January 1940 and run forward for 81 years through December 2020, spanning nearly the full period of ERA5 and SHIELD data. Although this period overlaps with the training data, which covers 1940–1995 and 2011–2019 (see “Methods”), ACE2 is only trained to predict two 6-hourly time steps ahead, and so the long autoregressive rollouts shown here demonstrate ACE2's ability to run stably and respond to long-term forcing. We

evaluate ACE2's inference performance on a held out 10-year test period in the subsequent subsections.

Figure 1 shows time series of global- and annual-mean variables for ACE2 and the reference datasets. Both ACE2-ERA5 and ACE2-SHIELD track the long-term trends of their reference datasets closely, which are driven largely by the forced SST trends. Differences in 2-meter air temperature between ERA5 and SHIELD themselves, despite the same SSTs, are largely from disagreement over high-elevation land and polar sea and land ice (not shown). Spatial patterns of long-term trends in the reference dataset are well-matched by ACE2 (Section S1.1 and Fig. S1). Shorter term inter-annual variability of 2-meter air temperature and total water path is also reflected in ACE2's predictions but is slightly muted compared to the reference datasets. The performance of ACE2 is similar between the training and validation periods and the held out test period (shaded light gray). In contrast, the previously trained ACE-climSST²¹ does not reproduce the historical moistening trends (Fig. 1c) when forced with AMIP SST; it also fails to predict historical warming in other temperature variables that is captured by ACE2 (not shown).

The ACE2-SHIELD and ACE2-ERA5 models chosen by our checkpoint selection criteria (best inference performance over 1940–2000, see Section “Checkpoint selection based on climate skill”) have similar skill in predicting inter-annual variability, comparable to the noise floor set by the SHIELD reference variability. Figure 1b, d shows a scalar skill metric (R^2) of the global- and annual mean series, including each of the four models in the training ensemble for ACE2-SHIELD. e.g., ACE2-ERA5 has a mean R^2 of 2-meter air temperature of 0.93, while for SHIELD reference variability the R^2 is 0.97. However, not all members of the training ensemble for ACE2-

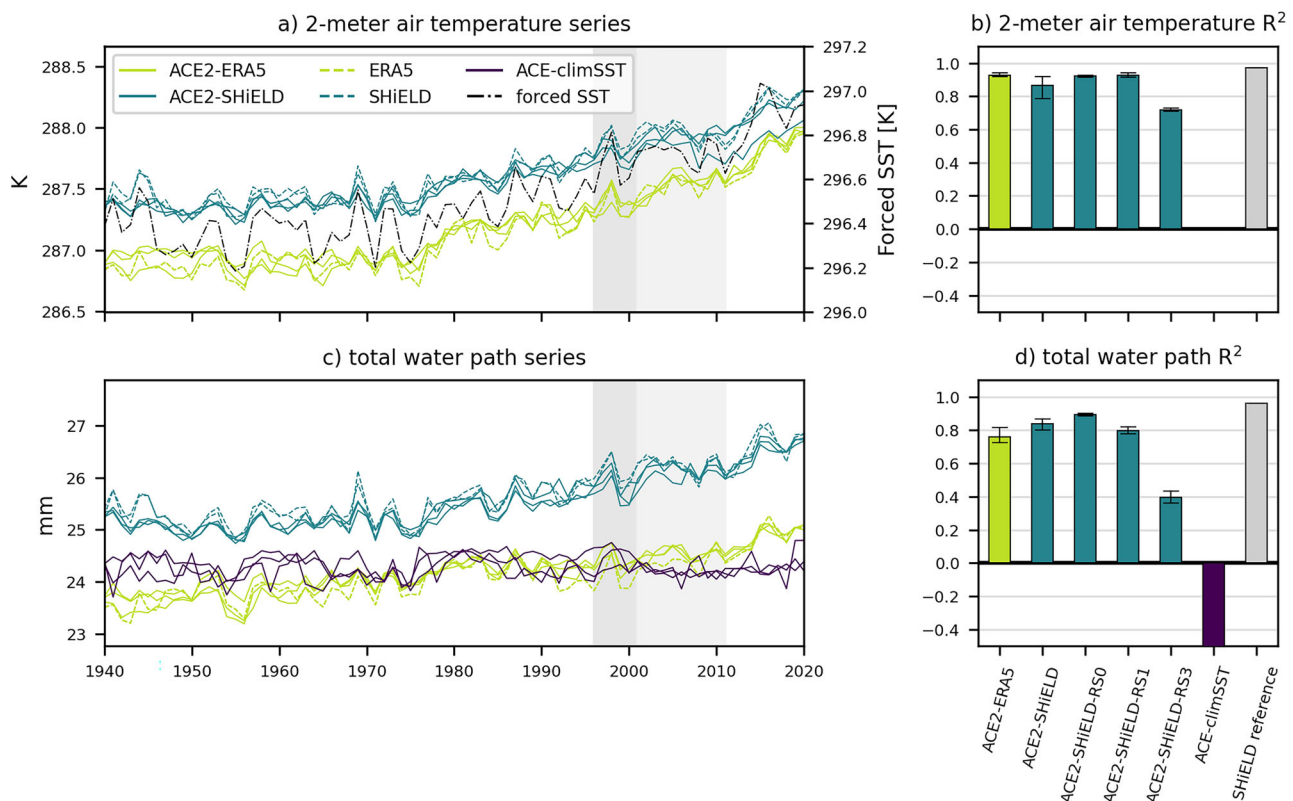


Fig. 1 | Global and annual-mean time series of ACE2 and reference datasets over 1940–2020. Global- and annual-mean series for a 2-meter air temperature and c total water path over 81-year evaluations of ACE2-ERA5 and ACE2-SHIELD. For each ACE2 evaluation, a three-member initial condition (IC) ensemble of the model (each initialized 1 day apart) is shown in solid lines, and the reference dataset is shown in dashed lines (e.g., ACE2-ERA5 vs. ERA5 itself). The validation and test periods are shaded in dark gray and light gray, respectively. As a baseline, the ACE-climSST model²¹ forced with the historical SST is also shown for total water path

(2-meter air temperature was not predicted by this model). The “forced SST” in a is the prescribed SST averaged over 45°S to 45°N in the SHIELD simulation (ERA5 SSTs are similar though not identical). The R^2 of the 81-year series are shown in (b) and (d). For ACE2-SHIELD, the skill metrics for each of four trained models are shown. Error bars indicate the range over three IC ensemble members for each model. The “SHIELD reference” bar in (b) and (d) is the R^2 computed between the two SHIELD ensemble members which differed only in initial condition.

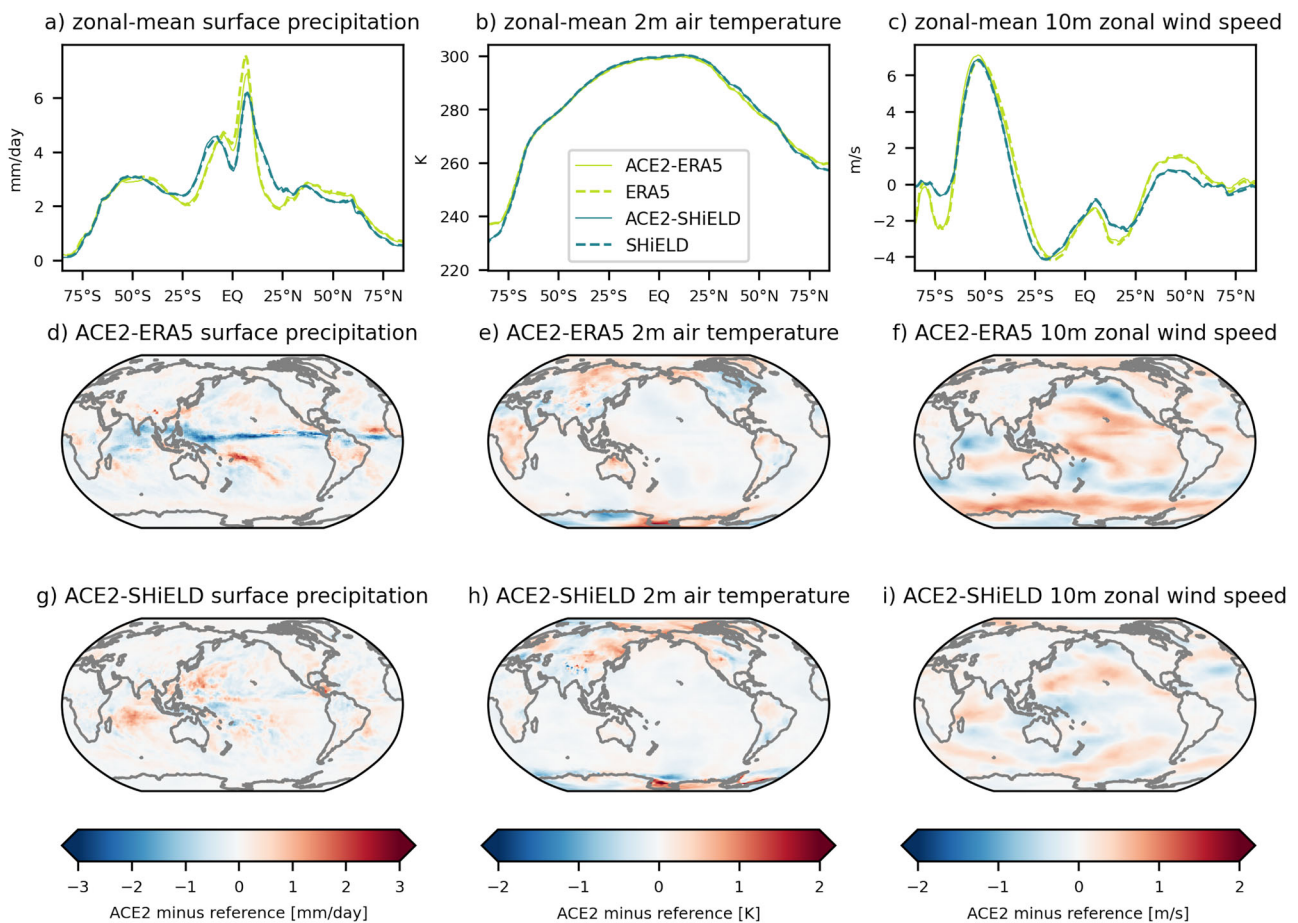


Fig. 2 | Time-mean climate biases of ACE2 over the 10-year test period.

a–c Zonal- and time-mean for ACE2 (solid) and its reference datasets (dashed) over test period spanning 2001-01-01 to 2010-12-31, for selected variables. **d–f** ACE2-

ERA5 time-mean biases over this time period. **g–i** ACE2-SHiELD time-mean biases over this time period. Results for a single initialization of each ACE2 model are shown.

SHiELD have the same skill; one of the trained models (labeled “-RS3”) has much poorer skill than the other three.

Climate skill: time-mean biases

We evaluate ACE2’s inference performance on a 10-year simulation forced by SSTs and CO₂ from the test period 2001-01-01 to 2010-12-31. Figure 2a–c shows the zonal- and time-mean of the ACE2-ERA5 and ACE2-SHiELD predictions. Each model’s predictions adhere closely to its reference dataset in zonal- and time-mean, such that ACE2 errors are much smaller in magnitude than the difference between the ERA5 and SHiELD datasets themselves.

The time-mean bias spatial patterns of ACE2-ERA5 and ACE2-SHiELD are different for surface precipitation and 10-meter wind speed (Fig. 2d, g, f, i), but for both models the largest precipitation errors are around the oceanic tropical convergence zones, where time-mean precipitation is large. The models’ bias patterns are more similar for 2-meter air temperature (Fig. 2e, h) with larger-magnitude temperature biases over high-latitude land and sea ice. Over ocean regions the temperature biases are smaller, as expected due to their strong coupling with the specified SST.

To quantify the magnitudes of the biases above, global time-mean RMSEs (Eq. (9)) of key surface fields over the 10-year test period are shown in Fig. 3. The errors of ACE2-ERA5 are computed with respect to the ERA5 dataset, while the ACE2-SHiELD and ACE-climSST errors are computed with respect to SHiELD. For all variables, the ACE2 models easily outperform the prior ACE model (ACE-climSST; ref. 27) and their errors are much smaller than the difference between the SHiELD and ERA5 datasets. To enable comparison with NeuralGCM, for which the time-mean error of

total water path over a 1-year simulation was reported (c.f. Fig. 4i of ref. 13) we run an analogous ACE2-ERA5 simulation spanning 2020, a period not used for training or validation. ACE2-ERA5 has similar error as NeuralGCM, about 1.05 mm versus 1.09 mm, respectively, over this period (Fig. 3c).

The error magnitudes of ACE2-ERA5 and ACE2-SHiELD against their reference datasets are similar for 2-meter air temperature. However, for surface precipitation rate and total water path, ACE2-ERA5 has a larger error than ACE2-SHiELD, possibly indicating a particular challenge in emulating moisture variables in the ERA5 dataset, which involves a data assimilation scheme and is natively at higher horizontal resolution than the SHiELD model. For ACE2-SHiELD, the error magnitudes are typically only 1.1–1.5 times the SHiELD reference variability (i.e., the magnitude of differences between the two SHiELD ensemble members, sampled over different 10-year periods). That is, by this metric, the 10-year mean climate of ACE2-SHiELD is nearly indistinguishable from that of the reference model.

The ACE2-SHiELD training ensemble shows non-trivial variability between models; the selected model (“ACE2-SHiELD”) slightly outperforms the other models (“ACE2-SHiELD-RS0”, “-RS1”, “-RS3”) over the test period (Fig. S9). See Section S1.5 on model selection for more information.

Atmospheric response to ENSO variability

We compute the atmospheric response to the El Niño-Southern Oscillation (ENSO; ref. 28) by regressing the predicted variables onto the Niño 3.4 index (see Eq. (13)). Maps of the ENSO-regressed surface precipitation rate for

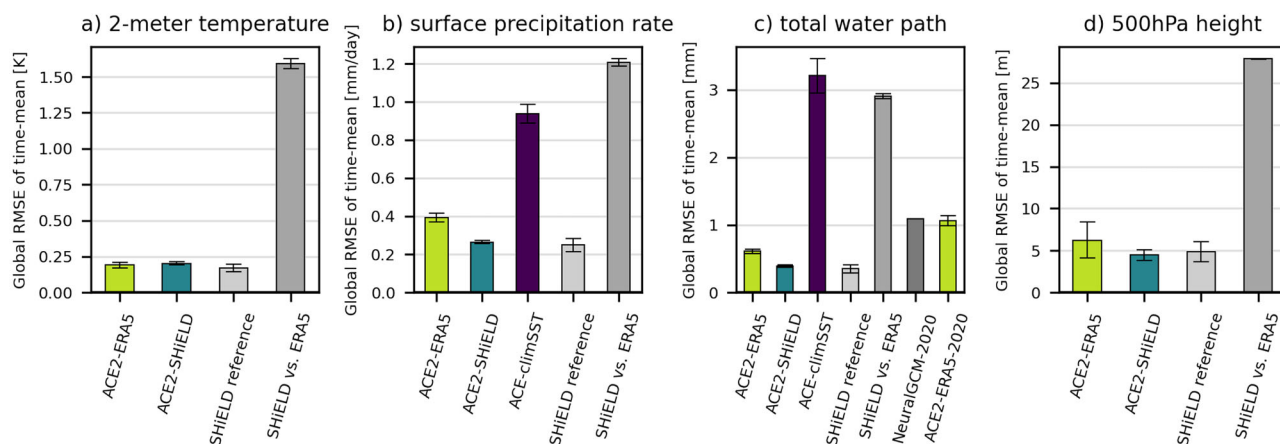


Fig. 3 | Global RMSE between the time-mean of ACE2 and its reference dataset (ERA5 or SHIELD). The global RMSE of the time-mean (Eq. (9)) for **a** 2-meter air temperature, **b** surface precipitation rate, **c** total water path and **d** 500 hPa geopotential height. Error bars indicate the 95% confidence interval based on the IC ensemble. Also included are NeuralGCM error against ERA5, SHIELD reference

variability, the error of ACE-climSST evaluated against the SHIELD dataset, and the error of the SHIELD simulations against ERA5. ACE-climSST did not predict 2-meter temperature or 500 hPa height. NeuralGCM¹³ results are only available for total water path for a single year (2020), and so we also show 2020-only results of ACE2-ERA5.

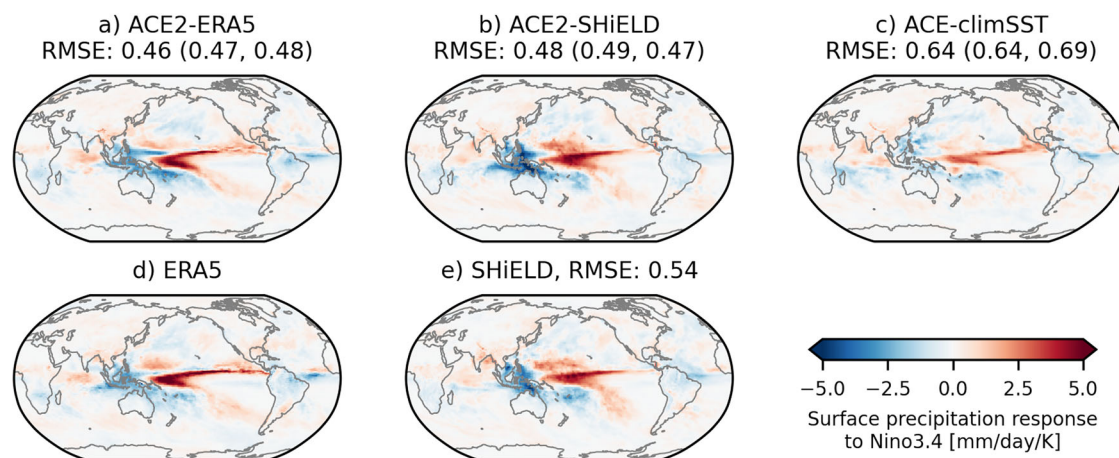


Fig. 4 | Maps of regression coefficients of predicted and reference dataset surface precipitation against the Niño 3.4 index over the 10-year test period. Single model initializations are shown. **a** ACE2-ERA5, **b** ACE2-SHIELD, **c** ACE-climSST evaluated on SHIELD, **d** ERA5 reference, **e** SHIELD reference, all for the 10-year test period. Titles of (a–c) indicate the RMSE of the predicted map

against its reference map; the numbers in parenthesis are for the two other initializations that are not shown. For **e**, the SHIELD reference variability is calculated as the RMSE between the regression coefficient maps of the two ensemble members.

ACE2-ERA5, ACE2-SHIELD, their reference datasets, and ACE-climSST are shown for the 10-year test period. ACE2 reliably reproduces the canonical response of surface precipitation to Niño 3.4 variability²⁸ in which positive Niño 3.4 is associated with increased precipitation in the central tropical Pacific and western Indian Ocean, and decreased precipitation over the maritime continent and tropical Atlantic (Fig. 4). Furthermore, ACE2 clearly reproduces the details of the Niño3.4 regression maps in the respective ERA5 and SHIELD reference datasets.

In contrast, the previous ACE-climSST predictions show a somewhat skillful but muted precipitation response to Niño 3.4 when evaluated using SHIELD forcing (Fig. 4c), demonstrating the value of ACE2 over ACE-climSST, most notably due to training on datasets with historical SST variability. The RMSE of the precipitation ENSO regression maps for ACE2-ERA5 (0.46 mm/day/K) and ACE2-SHIELD (0.48 mm/day/K) are smaller than that of ACE-climSST (mean 0.64 mm/day/K), and are comparable to the internal variability of this regression map in SHIELD (0.54 mm/day/K). A similar result is found for outgoing longwave radiation at top of atmosphere (Fig. S2). The extratropical teleconnections of the atmospheric circulation to ENSO over boreal winter are discussed in Supplementary Information

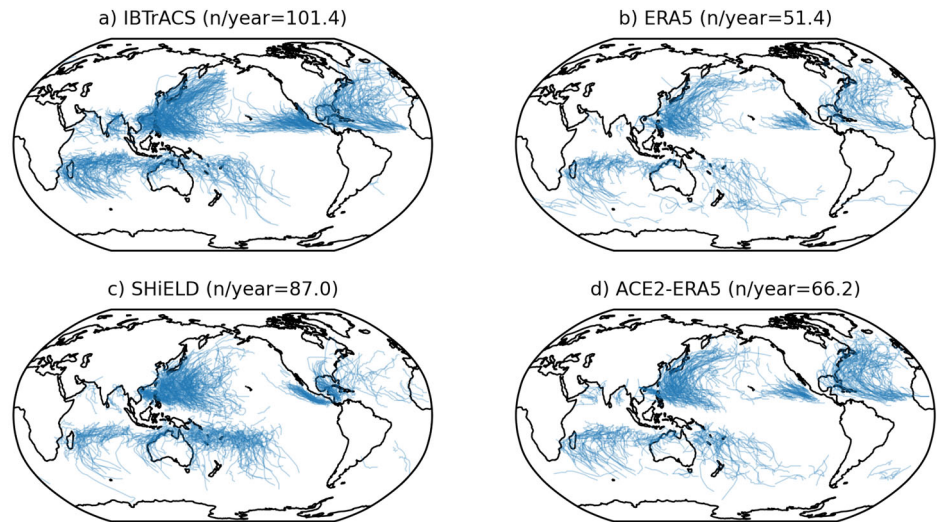
(Figs. S3 and S4). Briefly, ACE qualitatively shows the expected deepening of the Aleutian low and dipole of precipitation over North America.

Maps of ENSO coefficients for ACE2 rollouts spanning the entire 81-year ERA5/SHIELD period (see Figs. S3 and S4) are qualitatively similar to those for the 10-year test period, showing that the learned response to ENSO is robust.

Tropical cyclone climatology

Tropical cyclones are particularly damaging weather phenomena whose characteristics, such as strength and intensification rate, are projected to change with global warming^{29–31}. Their accurate representation would be a valuable feature of climate model emulators to allow the assessment of changes in these properties as a function of changing boundary conditions. In this section, we compare the strength, frequency, and location of tropical cyclone-like features in the ERA5 dataset, the ACE2-ERA5 emulator and, for comparison, the C96 (approximately 100 km resolution) SHIELD atmospheric model. However we note the SHIELD atmospheric model at C96 resolution is not expressly designed or intended to accurately represent tropical cyclones.

Fig. 5 | Climatology of tropical cyclone tracks over 10-year test period. Tracks of tropical cyclone-like features over the 2001–2010 period for **a** the IBTrACS dataset, **b** ERA5, **c** the C96 SHIELD model and **d** ACE2-ERA5. The tracks for (**b–d**) are determined based on minima in sea-level pressure along with maxima in upper troposphere temperature. See main text for details. The average number of tropical cyclones across the globe per year is shown in the title of each panel, although the IBTrACS dataset is not directly comparable to the detections in other panels which use a tracking algorithm applied to 1° resolution data.



The features are detected using 1° horizontal resolution data, although tropical cyclones are not well resolved at this horizontal resolution (e.g., their strength is often underestimated³² or they may be simply not detected). We use the TempestExtremes package and apply the default setting recommended for detecting tropical cyclones (Section 3.2 of ref. ³³), noting that these defaults were originally tuned for ERA5 at 0.25° resolution. One exception is that instead of using upper-level geopotential thickness (Z300 minus Z500) to detect warm cores aloft, we use upper tropospheric temperature since ACE2 does not directly predict geopotential height. Specifically, we use T_3 , which is the mean temperature between about 250 hPa and 400 hPa (see Tables S1 and S2). Instead of requiring a thickness decrease away from the tropical cyclone center, we require a temperature decrease of 0.4 K. Assuming hydrostatic balance, this is approximately equal to the $58.8 \text{ m}^2 \text{ s}^{-2}$ thickness decrease suggested in ref. ³³.

Figure 5 shows the tropical cyclone tracks detected for ERA5, ACE2-ERA5 and SHIELD according to the above criteria as well as those in the IBTrACS database^{34,35} for the 10-year test period (2001–2010). The number of cyclones detected per year globally is shown in the title of each panel, although we note that this quantity is sensitive to the parameters chosen for the detection algorithm used in Fig. 5b–d. That said, since the same detection parameters are used for ERA5, ACE2-ERA5 and SHIELD, we can compare this quantity between these datasets. Globally, ACE2-ERA5 overpredicts tropical cyclone frequency by about 28% compared to its target dataset ERA5. The SHIELD atmospheric model predicts about 69% more tropical cyclones than ERA5 at the given resolution, which may be more in line with the true global frequency of tropical cyclones (Fig. 5a). Regionally, ACE2-ERA5 closely matches the basin-by-basin frequency of tropical cyclones in the ERA5 dataset (Fig. 5). On the other hand, compared to ERA5 and IBTrACS, the SHIELD atmospheric model has too few tropical cyclones in the North Atlantic. Overall, this analysis suggests the ACE2-ERA5 emulator accurately captures the regional frequency of tropical cyclone-like events in the ERA5 dataset.

A possible concern with our evaluation framework is that ACE2-ERA5 is forced with observed sea surface temperatures that contain a signature of past tropical cyclones, which can leave behind a cold wake³⁶. Hypothetically, the machine learning emulator could learn to generate tropical cyclones based on the prescribed sea surface temperature signature. However, when we force ACE2-ERA5 with a climatological sea surface temperature dataset, we recover a very similar frequency and distribution of tropical cyclones as when we force it with historical sea surface temperature, showing that this is not the case (Fig. S6).

The strength of the detected tropical cyclones, as measured by minimum sea level pressure and maximum 10 m wind speed, is also accurately

emulated by ACE2-ERA5 (Fig. S7) when compared to the ERA5 dataset. The SHIELD model tends to produce more cyclones with strong ($>30 \text{ m/s}$) near-surface wind speeds.

Tropical precipitation variability

Prior work has confirmed that ACE is able to closely replicate the precipitation variability in a coarse resolution atmospheric model²². Here we show a brief analysis of tropical precipitation variability focused on ACE2-ERA5 since the ERA5 dataset contains variability, such as equatorial Kelvin waves or the Madden-Julian Oscillation, which is often missing or too weak in coarse resolution atmospheric models (c.f. Fig. 17d of refs. ^{37,38}).

Figure 6 shows the tropical-mean precipitation over longitude and time for the 2007–2008 period, which contained several strong Madden-Julian Oscillation (MJO) events in the observed record³⁹ that are apparent in the ERA5 dataset, for example during December 2007 (Fig. 6a). The shown ACE2-ERA5 and SHIELD simulations (Fig. 6b, c) are initialized in 2001 and 1939 respectively, so we do not expect the timing of events to coincide between the three datasets. However, it is notable that the spatio-temporal variability of the ERA5 dataset is much more closely captured by ACE2-ERA5 than it is by SHIELD. For example, relatively small-scale eastward propagating Kelvin waves⁴⁰ exist in both the ERA5 and ACE2-ERA5 precipitation variability, but are less apparent in SHIELD. ACE2-ERA5 does show some notable differences from ERA5, for example generally being smoother in longitude and time.

To more explicitly compare the representation of the MJO, the dominant mode of intraseasonal variability in the tropics⁴¹, we compute a lag-correlation diagnostic which demonstrates the eastward movement of precipitation on the MJO timescale (20–100 days) around the Indian Ocean and Maritime Continent⁴². Specifically, we first compute $P_{10^\circ\text{S}-10^\circ\text{N}}^{20-100\text{day}}$, which is the surface precipitation rate averaged between 10°S and 10°N and bandpass filtered between 20 and 100-day variability. We then compute the lag correlation of $P_{10^\circ\text{S}-10^\circ\text{N}}^{20-100\text{day}}$ at all longitudes with that over the western Indian Ocean (80°E and 100°E). Figure 7 shows this lag-correlation for ERA5, ACE2-ERA5 and SHIELD, in all cases computed over 2001–2010. This demonstrates the eastward propagation of the MJO in ERA5 (Fig. 7a) while the SHIELD model lacks coherent eastward propagation of precipitation variability in this region (Fig. 7c), a fairly common and longstanding issue of coarse resolution global atmospheric models^{38,43,44}. However, ACE2-ERA5 shows an eastward propagation of the MJO consistent with ERA5, both in terms of phase speed and longitudinal extent. This lends further credibility to the realism of ACE2-ERA5 emulator's representation of tropical variability on sub-seasonal timescales.

Fig. 6 | Hovmöller plot of tropical precipitation in ACE2-ERA5 and the reference datasets. Daily-mean precipitation rate averaged between 10°S and 10°N over the 2007–2008 period for **a** ERA5 and **b** the 10-year ACE2-ERA5 run initialized on 2001-01-01 and **c** the first ensemble member of the SHIELD AMIP simulation.

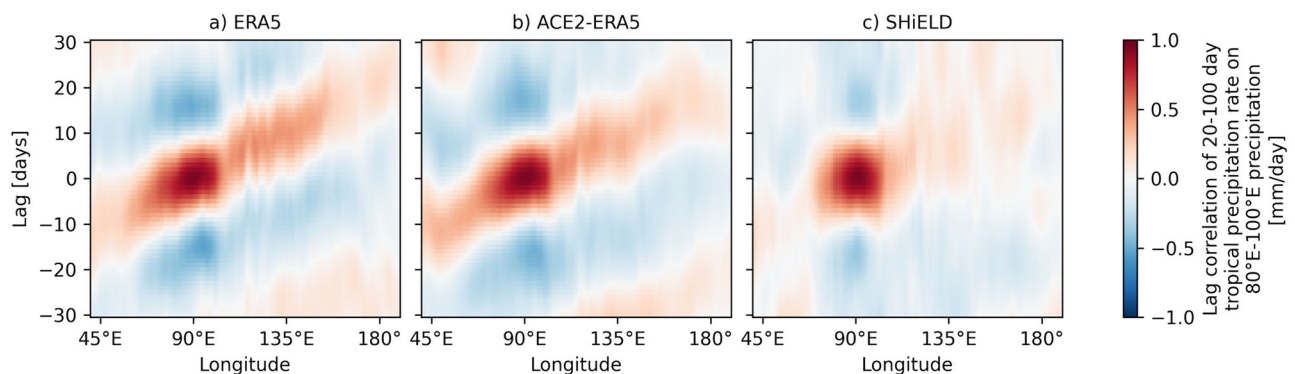
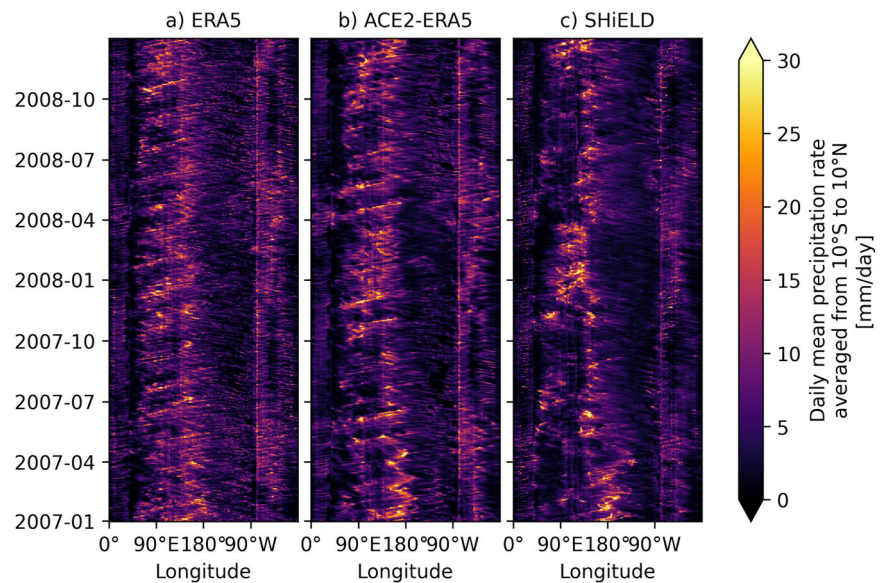


Fig. 7 | Lag correlation of $P_{10^{\circ}\text{S}-10^{\circ}\text{N}}^{20-100\text{day}}$ at all longitudes with $P_{10^{\circ}\text{S}-10^{\circ}\text{N}}^{20-100\text{day}}$ averaged from 80°E to 100°E (ref. 42). $P_{10^{\circ}\text{S}-10^{\circ}\text{N}}^{20-100\text{day}}$ is the surface precipitation rate averaged from 10°S to 10°N, and then filtered with a 20–100 day bandpass filter. Calculated

over the 2001–2010 period from **a** ERA5, **b** the test period run of ACE2-ERA5 and **c** the first ensemble member of the SHIELD AMIP run.

Polar stratospheric variability

Existing machine learning models for weather prediction either do not explicitly resolve the stratosphere^{23,45}, do not report on skill in the stratosphere^{9,46}, or show relatively worse short-term predictive performance in the stratosphere compared to lower vertical levels¹⁰. The uppermost vertical layer of ACE2 represents a mass-weighted integral of atmospheric properties (temperature, horizontal winds and moisture) between approximately 50 hPa and the top of atmosphere (Table S2). Therefore, we are able to evaluate the representation of large-scale stratospheric processes. In this section, we focus on comparing polar stratospheric variability in ERA5 and ACE2-ERA5. The variability in the strength of the stratospheric polar vortex—as measured by the zonal mean wind u_0 in ACE2’s top vertical layer at 60° latitude—is the dominant mode of sub-seasonal variability in the stratosphere. It is an important source of sub-seasonal to seasonal predictability⁴⁷ and is a strong control on ozone chemistry, resulting in the ozone hole being most evident in the Southern Hemisphere⁴⁸.

ACE2-ERA5 reproduces the expected seasonal asymmetry in mean polar stratospheric vortex strength and variability (Fig. 8). By overlaying the zonal mean u_0 at 60°N and 60°S for each of the 10 years from the test period, we see the expected variability in the Northern Hemisphere exists in ACE2-ERA5. This includes sudden stratospheric warming events in which the strength of the vortex rapidly decreases and the zonal-mean flow reverses. As expected, in the Southern Hemisphere, the average winds are

stronger while also being less variable from year to year. With only ten years for comparison, it is difficult to quantitatively compare the statistics of variability between ERA5 and ACE2-ERA5, but the qualitative behavior shown in Fig. 8 is promising. Longer simulations, which overlap with the training and validation periods, demonstrate good agreement between the 5th and 95th percentiles of u_0 at 60°S and 60°N (not shown).

While ACE2-ERA5 shows some variability of near-equatorial stratospheric winds between eastward and westward with approximately the same magnitude as the observed quasi-biennial oscillation⁴⁹, the variability is irregular and does not have the correct period (not shown).

Weather skill

Although accurate weather forecast skill was not a primary objective of this work, in this section we assess ACE2-ERA5’s medium range global forecast skill. Figure 9 shows global RMSE averaged over 10-day forecasts initialized throughout the 2020 period for T_{2m} , T_{850} , Z_{500} and v_{10m} (see Table S1 for definitions). As baselines, we use Graphcast¹⁰ and the “ERA5 forecasts” as provided by WeatherBench 2.0⁵⁰, both compared with the ERA5 dataset. The “ERA5 forecasts” are forecasts using ECMWF’s IFS model, with the same model version used to produce the ERA5 reanalysis and initialized from ERA5 snapshots to provide a more direct comparison with models such as ACE2-ERA5. The ACE2-ERA5 forecasts in Fig. 9 correspond to 48 initializations equally spaced across 2020, while Graphcast and era5-

Fig. 8 | Annual cycle of zonal-mean u_0 (eastward wind vertically integrated from ~50 hPa to top of atmosphere) at (top row) 60°N and (bottom row) 60°S for (left column) ERA5 and (right column) ACE2-ERA5. For ERA5, each of the years from 2001–2010 test period are plotted. For ACE2-ERA5, a simulation is initialized from ERA5 on 2001-01-01 and run for 10 years. Individual gray lines show each year, while the bold black line shows the average over the 10-year period.

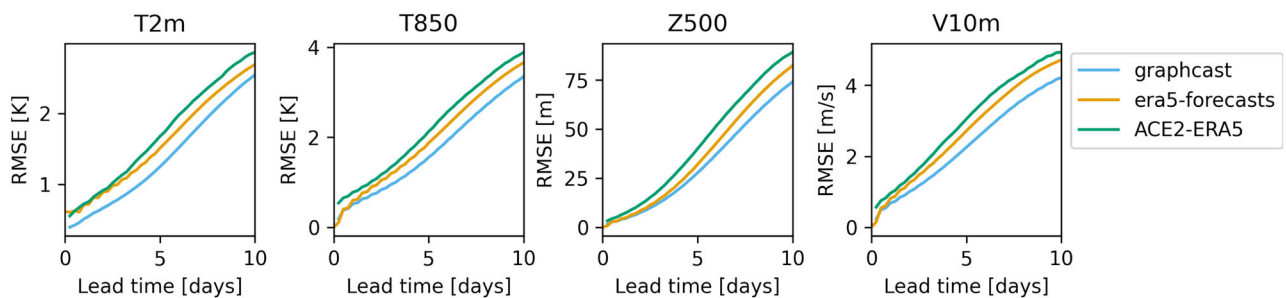
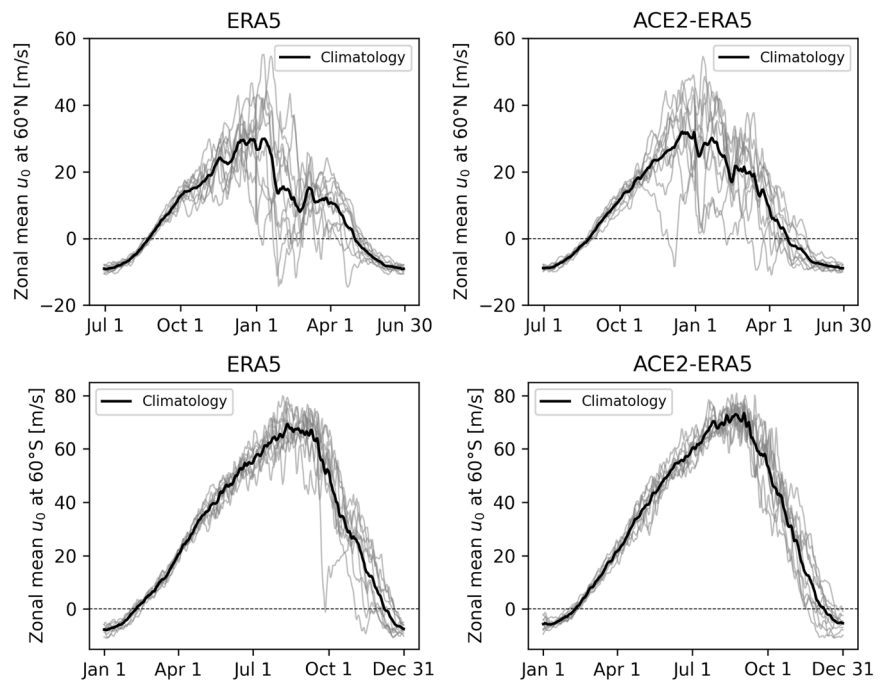


Fig. 9 | RMSE of ACE2-ERA5 during 2020, compared to GraphCast and IFS initialized from ERA5 (“era5-forecasts”). In order to make a fair comparison, for these ACE2-ERA5 simulations the sea surface temperature and sea-ice fraction are kept fixed at their initial values instead of being prescribed throughout the simulation.

forecasts consist of forecasts initialized at 0Z and 12Z on every day of 2020. Figure 9 shows that ACE2-ERA5 is slightly behind the ERA5-version of IFS (e.g., half a day at 5-day lead time for 850 temperature) and further behind Graphcast by another day. Although there are a variety of differences between ACE2-ERA5 and Graphcast, we believe the most substantial to be (1) the different vertical coordinate (ACE2-ERA5 uses a terrain-following coordinate, Graphcast a pressure coordinate) and (2) the architecture underlying each model (SFNO and Graph Neural Network respectively). Preliminary work has found that using a different architecture but keeping ACE2’s terrain-following vertical coordinate can lead to a significant increase in weather forecast skill, but not mean climate skill (not shown), suggesting that the architecture difference is likely the more important one.

Millennial timescale stability

To test the stability of ACE2 models over a longer duration than the length of the AMIP forcing dataset (about 80 years), we generate a climatological forcing dataset which can be repeated indefinitely. This is computed by averaging surface temperature, CO_2 and sea-ice fraction from ERA5 over the 1990–2020 period, resulting in a 6-hourly climatology estimate. We then initialize an ACE2-ERA5 simulation from ERA5 on 2001-01-01 and run it for 1000 years, forced by the annually repeating climatological dataset. No signs of instability (i.e., indefinitely growing errors) are seen in this 1000-year run, and the time-mean climate is nearly identical between different 100-year periods of the simulations. As an example, Fig. 10 shows the global-

mean total water path timeseries for the first 100 and last 100 years of the simulation. There is no long term drift in total atmospheric moisture, and the seasonal cycle remains of consistent amplitude throughout the simulation. This is a noteworthy improvement on ACE-climSST, where 100-year simulations showed unrealistic fluctuations in the amplitude of the global-mean seasonal cycle (c.f. Fig. 10 of ref. 21).

Learning at coarser horizontal resolution

Traditional climate models often achieve improved skill at increasing resolution, as physical processes are more accurately represented. However, this is not necessarily the case for coarse emulators of a climate model without an explicit representation of atmospheric processes. Here we compare the performance of ACE2 trained on the SHIELD AMIP dataset coarsened to 4-degree resolution against the coarsened output of ACE2 trained at 1-degree resolution (as presented in Section “Results”). Ideally, the climate of the 4-degree ACE2 emulator could be just as skillful as that of the 1-degree emulator, but is this achievable in practice?

With identical training and inference regimes, the time-mean ensemble-mean biases of 2 m air temperature (T_{2m}) and precipitation have slightly higher magnitudes for the models trained at 4° resolution compared with 1° resolution (Fig. 11). Both have biases much smaller than a C24 (approximately 400 km resolution) SHIELD baseline simulation. The largest T_{2m} biases are at high latitudes. T_{2m} biases over open ocean regions are minimal, as physically expected due to strong coupling of T_{2m} with SST.

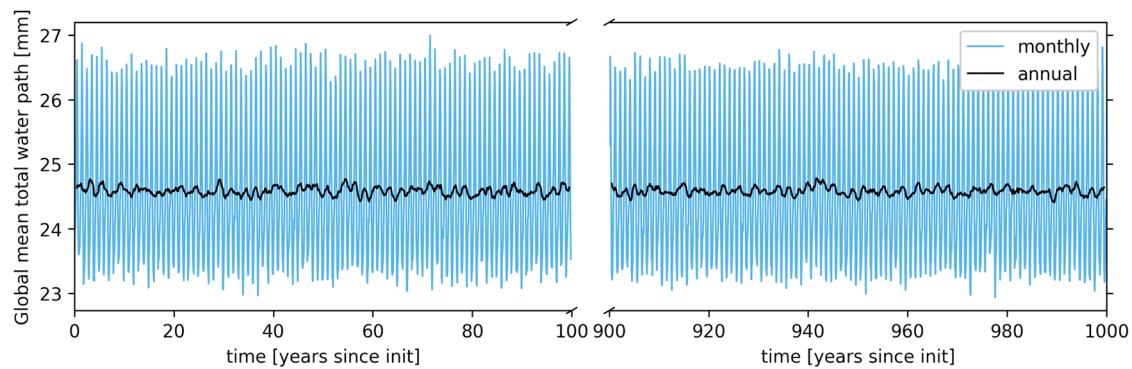


Fig. 10 | The global mean total water path for the first and last 100 years of a 1000-year long simulation with ACE2-ERA5 forced with 1990–2020 climatological mean sea surface temperatures, land type fractions and CO₂. Shown for (blue) monthly mean and (black) annual mean.

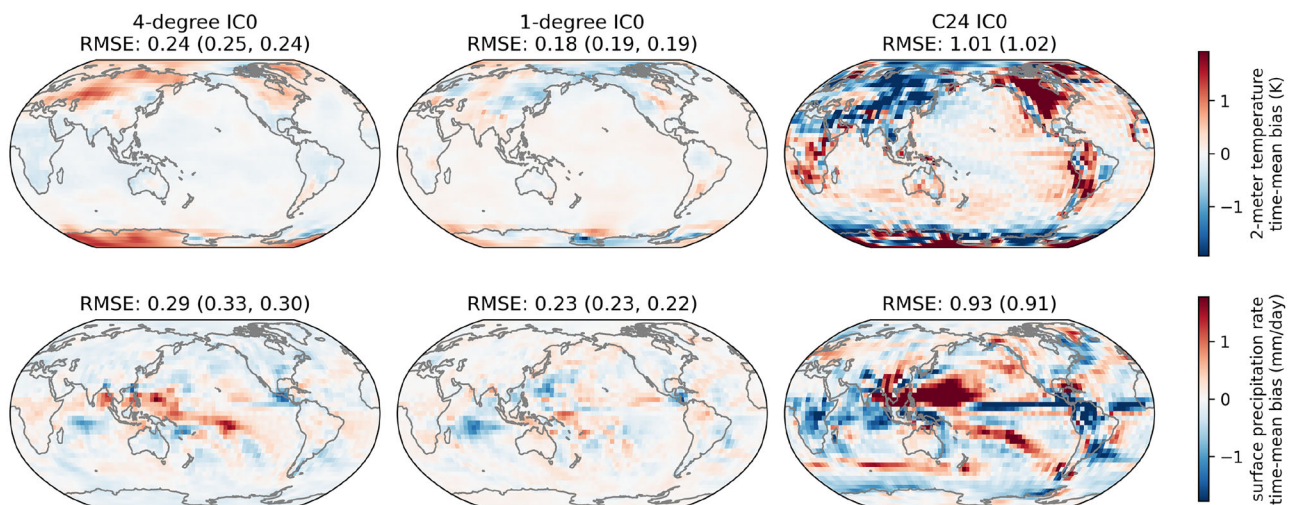


Fig. 11 | Single initial condition time-mean biases of T_{2m} and precipitation for 10-year inference using 1° and 4° ACE2-SHiELD models and the C24 (4°) SHiELD baseline model, with respect to the C96 (1°) SHiELD model. 1° values are area-weighted block-coarsened by a factor of 4 prior to computing RMSE. Values are

shown for the same time period and ensemble configuration as in Fig. 4. RMSE is shown for the ensemble member shown in the map, with values for the other two members shown in parentheses.

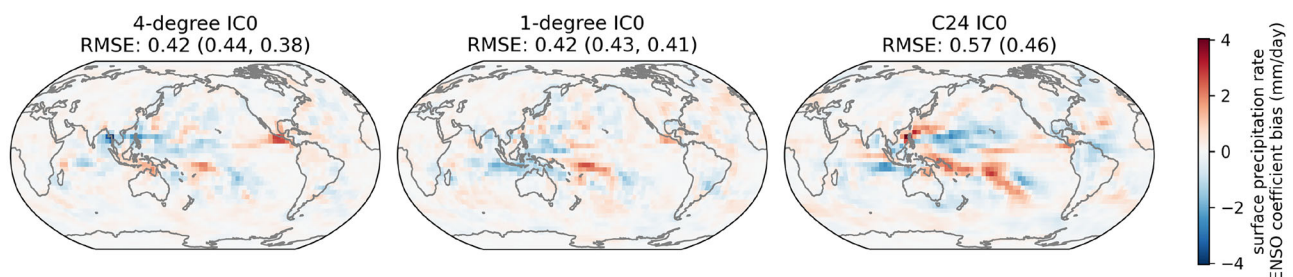


Fig. 12 | Bias of single initial condition ENSO regression coefficient maps of surface precipitation rate for 3-member 10-year inference using 1 and 4-degree ACE2 models and a C24 (4°) SHiELD baseline model, with respect to the C96 SHiELD model. 1-degree values are area-weighted block-coarsened by a factor of 4.

Values are shown for the same time period and ensemble configuration as in Fig. 4. RMSE is shown for the ensemble member shown in the map, with values for the other two members shown in parentheses.

The biases in time-mean precipitation are largest at low latitudes, in regions of large mean precipitation. We use the C24 SHiELD as a baseline because coarsening spatial resolution is a common strategy to decrease the computational cost of physics-based atmospheric models. However, ACE2 is still about 25x more energy efficient than C24 SHiELD and it is about 700x more energy efficient than C96 SHiELD (Table 3).

The patterns of precipitation variability regressed on ENSO variability have similar RMSE amplitudes for the 4° emulator and the 1° emulator, and their biases with respect to the C96 SHiELD model

share many of the same spatial structures (Fig. 12). Both have lower biases with respect to the C96 SHiELD model than the C24 SHiELD baseline.

This similarity in skill between ACE2 trained at 1° and 4° is encouraging because it suggests that, unlike for physics-based climate models, a computationally light coarse emulator that might be attractive for paleo-climate or marine biogeochemistry applications can simulate coarse-scale climate features almost as well as a more expensive, memory-intensive fine-grid emulator.

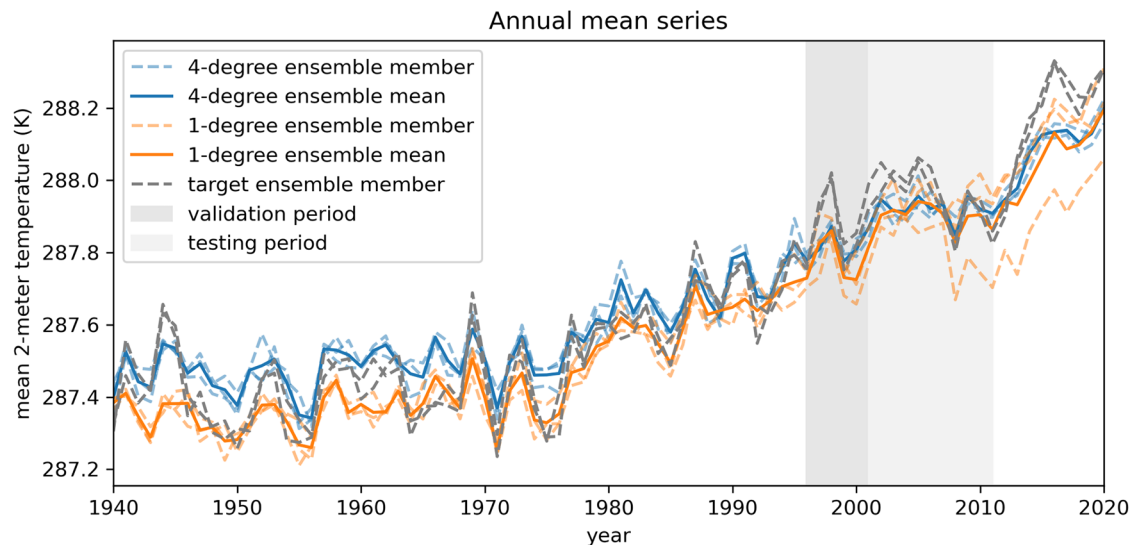


Fig. 13 | Annual and global mean 2-meter temperature for 81-year inference using 3-member initial condition ensembles of 1 and 4-degree ACE2-SHiELD models. Values are shown for the same time period and ensemble configuration as shown in Fig. 1.

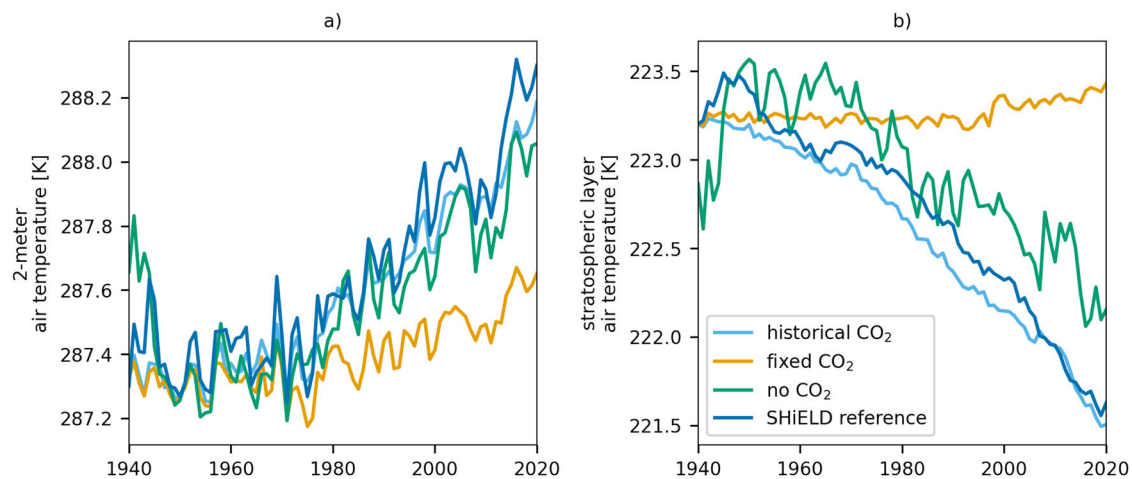


Fig. 14 | Global and annual-mean time series of 2 m and stratospheric air temperature for ACE2-SHiELD simulations with differing CO₂ inputs. Global- and annual-mean **a** 2-meter air temperature and **b** level-0 (stratospheric) air temperature. Shown for the standard ACE2-SHiELD model (“historical CO₂”); the same model but

with CO₂ concentrations fixed at the 1940 value (“fixed CO₂”); a version of ACE2-SHiELD trained without CO₂ as a feature (“no CO₂”); and the SHiELD reference data with historical SSTs and CO₂ forcing. The average of 3-member IC ensembles are shown.

The 1° and 4° models show a similar ability to reproduce the long-term trend and interannual variability (Fig. 13). Both models show reduced interannual variability over shorter timescales. Over the 1940–1975 period the 1° model is biased low and the 4° model is biased high. This leads to a better representation of the overall trend in the 1° model, as both models are biased low during 1996–2020.

We would also note that while we have trained the 4° model with the same hyperparameters as the 1° model for consistency, coarse model performance benefits from a larger embedding dimension, likely due to the increased subgrid activity at coarser resolution.

Using CO₂ as an input feature

ACE2 uses both global-mean CO₂ concentration and spatially-varying SST as forcing when trained on either ERA5 or SHiELD. During the AMIP period, historical global-mean SSTs and CO₂ both increase with time, and the physical causality (i.e., gradual uptake of heat by the oceans due to increased radiative heating from elevated CO₂) may be difficult to learn from 6-hourly changes in the atmospheric and SST states. Here we evaluate

the sensitivities of ACE2 to CO₂ specifically, by comparing ACE2 simulations with historical CO₂ to those where we set the concentration to a fixed value (1940 concentration of 307ppm), while retaining increasing SSTs.

Figure 14 shows that both near-surface and stratospheric global-mean temperature series in ACE2-SHiELD approximately match those in the reference dataset, when ACE2-SHiELD is forced with both historical SSTs and CO₂. There is both near-surface warming with polar amplification (Section S1.1) and near-uniform stratospheric cooling. When holding CO₂ fixed, ACE2 no longer produces stratospheric cooling, as is expected physically⁵¹. However, it also loses much of the trend of near-surface warming, which is not expected. This is largely due to lack of warming over high-latitude land (not shown), despite evidence that such polar amplification should be driven largely by SST and sea ice coverage forcing⁵².

In contrast, a version of ACE2 trained with SSTs but not CO₂ as forcing has global trends of near-surface warming and stratospheric cooling that somewhat underestimates these trends in the reference data, as well as excess inter-annual variability of stratospheric temperature. Thus using CO₂ as forcing with AMIP training datasets appears to improve the representation

of some aspects of CO₂-induced trends, while introducing non-physical relationships in others. We lack SHIELD simulations forced by historical SSTs and fixed CO₂ (and vice versa), but these could be generated to augment ACE2's training data and test whether this improves the physical sensitivities of the emulator.

Discussion

This study demonstrates the feasibility of training a machine learning emulator to accurately generate atmospheric variability and forced responses from time scales of days to decades. ACE2 has a realistic global mean atmospheric response to increased sea surface temperature and CO₂. It generates realistic variability including the atmospheric response to El Niño, the Madden Julian Oscillation, the geographic distribution of tropical cyclones and stratospheric polar vortex strength variability. By formulating ACE2 as an autoregressive model which simulates century-long trends through stepping forward 6 h at a time, we can ensure physical consistency. Specifically, ACE2 exactly conserves dry air mass and moisture. Furthermore, by simulating climate as the average of explicitly resolved weather, interpretability is improved. As an example, the mechanisms by which ACE2 simulates the correct atmospheric response to El Niño could be explored in a manner analogous to traditional numerical models.

The enormous decrease in computational expense for running ACE2 compared to a similar resolution physics based model (about 750x less energy cost and 130x faster throughput) opens the door to new use cases. With proper validation, generation of very large ensembles will allow accurate estimation of internal variability and the likelihood of extreme events, with important applications for climate risk estimation. The relative ease of running ACE2 compared to conventional global atmospheric models—it is possible with reasonable throughput even on a modern Apple laptop—allows a wider range of users to directly apply climate models to their questions of interest. Indeed, providing a deep learning emulator with new releases of physics based models may become standard practice.

Limitations of this work include the particular datasets used. For example, due to training on data corresponding to the last 80 years, we do not expect ACE2 to be able to properly simulate the response to strong climate change (e.g., a doubling of CO₂ or significant increase in sea surface temperature). Indeed, preliminary experiments indicate unrealistic patterns of near-surface air temperature response, even to relatively small magnitude uniform SST perturbations (Section S1.3) with unexpected cooling over some land and sea-ice regions and insufficient warming over ocean.

Furthermore, the SHIELD and ERA5 datasets both have shortcomings in accurately representing the true past conditions of the atmosphere. SHIELD is a coarse atmospheric model, and has biases in its global circulation. While the ERA5 dataset involves a data assimilation scheme to constrain its state to remain close to observations, fields such as the surface precipitation rate and radiative fluxes are not constrained and exhibit non-trivial biases with respect to satellite and station observations^{17,53}. In addition, ACE2 itself does not accurately represent the expected atmospheric response to increasing sea surface temperature while keeping CO₂ fixed (Section “Using CO₂ as an input feature”) suggesting a need to encode the impacts of greenhouse gases in a more sophisticated manner. Training on exclusively historical data, a period over which forcings such as CO₂, methane, aerosols and insolation varied simultaneously, makes it difficult for a global machine learning model to disentangle the effects of these

individual terms. Finally, ACE2 does not exactly conserve global atmospheric energy, because it has a more complex budget equation than dry air mass or moisture and has significant non-conservation errors in atmospheric models such as SHIELD.

Concurrent work has trained ACE2 on SHIELD simulations spanning a wider range of CO₂ concentrations²⁰, demonstrating excellent skill for ACE2 in interpolating between idealized climate scenarios. How best to combine reanalysis data such as ERA5 with simulation data from future or past climates for training remains an important open question. In addition, the ability to simulate additional components of the climate system, such as ocean and sea ice, is a basic requirement for a useful climate model emulator. We expect progress to continue advancing quickly in the application of machine learning for climate modeling.

Methods

Versioning nomenclature

We use the following nomenclature to distinguish between versions of the ACE model. ACE-climSST refers to the first version of ACE²¹ which was trained on a dataset produced by forcing an atmospheric model with annually-repeating climatological SSTs and otherwise fixed external forcing. In this study, we introduce ACE2, which has an increased parameter count and updated loss function, introduces hard physical constraints on mass and moisture and uses a new checkpoint selection strategy in training, among other changes described below. We present results from training ACE2 on two distinct datasets, described in the next section. To distinguish these models, we will describe them as ACE2-SHIELD and ACE2-ERA5 respectively.

Datasets

Two datasets are used as targets for emulation (Table 1). The first is output from the SHIELD atmospheric model¹⁹ at C96 (approximately 100 km) resolution forced by observed sea surface temperatures and greenhouse gases from the 1940–2021 period. The latter is the ERA5 reanalysis dataset¹⁷ from 1940–2022. Other than their sources, the datasets are the same in terms of variable set (see Table S1) and resolution. ACE2, like ACE, combines the model-level fields for air temperature, specific total water and horizontal winds into eight vertical layers. The 2D prognostic variables are surface pressure, surface temperature over land and sea-ice, 2-meter air temperature and specific humidity and 10-meter horizontal winds. These latter near-surface variables are new additions compared to ACE²¹ and are included due to their human impact relevance and importance for ocean coupling. Additional variables, used as diagnostics (outputs) only are the top-of-atmosphere and surface radiative fluxes, surface latent and sensible heat fluxes, surface precipitation rate and, for convenience, the 500 hPa geopotential height and 850 hPa air temperature. Finally, forcing variables (i.e., inputs only) are sea surface temperature, global-mean carbon dioxide (broadcast to a spatially uniform global field), incoming solar radiation at the top of atmosphere, land fraction, ocean fraction, sea ice fraction and surface topography. The use of carbon dioxide as a forcing input is a change from ref. 21.

The reference data is horizontally interpolated to the 1° Gaussian grid and the 6 h temporal resolution used by ACE2 and ACE. All flux variables (e.g., radiative fluxes, precipitation) are time-averaged over the 6-h intervals in order to enable exact evaluation of atmospheric budgets at the 6-hourly time resolution.

Table 1 | Datasets used in this study

Name	Train period	Validation period	Test period
ERA5	1940–1995, 2011–2019, 2021–2022	1996–2000	2001–2010, 2020
SHIELD	1940–1995, 2011–2021	1996–2000	2001–2010

ERA5 is a reanalysis product, here coarsened to 1° horizontal resolution¹⁷. SHIELD is an approximately 100 km resolution global atmospheric model which was forced by historical sea surface temperatures¹⁹. For the SHIELD dataset, data are available from two ensemble members initialized from slightly different initial conditions on October 1, 1939, doubling the number of samples available.

SHIELD dataset

To generate multiple physics-based realizations of climate forced by historically observed sea surface temperatures, sea ice, and carbon dioxide, we make use of the public version of the SHIELD model developed at the Geophysical Fluid Dynamics Laboratory (GFDL)¹⁹. This is GFDL's developmental version of the FV3GFS model used in ref. 21. The two models share a significant fraction of their code, the most notable difference being that SHIELD computes all microphysical updates every vertical remapping timestep within the dynamical core, rather than splitting the microphysical updates between the dynamical core and the physics^{19,54}.

We run SHIELD at two horizontal resolutions, C96 (roughly 100 km) and C24 (roughly 400 km), with 79 vertical levels; C96 simulation output forms the basis of our target dataset, while C24 simulation output serves as a computationally inexpensive physics-based baseline. Other than those related to horizontal resolution and convection—here we use the latest versions of both the shallow and deep convection schemes—we configure the parameters of the model following how they were configured in the C3072 (roughly 3 km) resolution X-SHIELD runs of ref. 55. Note that no special tuning was attempted to help the climate of SHIELD better match observations when run at coarser resolution. However we reduced a parameter controlling the strength of the mountain blocking scheme in the C24 configuration to help its climate, particularly the near-surface temperature over land, better match that of the C96 configuration based on the scheme's empirical sensitivity to resolution (J. Alpert and F. Yang, personal communication, August 9, 2019).

At each horizontal resolution, we run two identically forced simulations over 1940–2021, but with different initial conditions. The initial conditions are generated by running a spin up simulation starting from GFS analysis for 2020-01-01 with 1930-01-01 forcing data for 117 months to 1939-10-01, outputting daily restart files from the last month. This roughly 10-year period is meant to allow the model to adjust to the historical forcing after being initialized with present-day atmospheric conditions; the time-scale is mainly limited by the time it takes stratospheric water vapor to equilibrate. The restart files from 1939-09-30 and 1939-10-01 represent the state with which we start the two ensemble members on 1939-10-01, providing 3 months of spin up time prior to 1940-01-01 to allow the model states to meteorologically diverge. A similar approach was used to generate initial conditions in the coupled model ensemble context in⁵⁶. We run the simulations until 2021-12-16T12:00:00, the last available time in our reference SST and sea ice dataset.

The historical SST and sea ice concentration data come from that used to force historical AMIP CMIP6 simulations^{18,57,58} and are provided on a 1° regular latitude-longitude grid as a monthly time series; space and time interpolation occurs online at the time of prescription within SHIELD. We prescribe carbon dioxide as a time series of annual and global means, with data prior to 2015 coming from that used for CMIP6⁵⁹ and data after coming from the NOAA Global Monitoring Laboratory⁶⁰; in these runs we assume CO₂ is well-mixed (i.e., globally uniform).

Data from these simulations is output on the model native cubed-sphere grid at 6-hourly intervals. We make use of GFDL's `fregrid` tool⁶¹ to conservatively regrid the model state to a Gaussian grid. In the case of C96 data this is a 1° grid, and in the case of C24 data this is a 4° grid. Similar to a regular latitude-longitude grid, a Gaussian grid provides increased resolution in the polar regions, which means that with a conservative regridding approach the original cubed-sphere grid cell edges in these regions are resolved with high fidelity. As in ref. 21, we perform a spherical harmonic transform (SHT) round trip on all but the sea-ice fraction, ocean fraction and land fraction variables in the regridded output to smooth these sharp boundaries, which otherwise produce artifacts under spherical harmonic transforms. Finally we coarsen vertically resolved fields from the native 79 vertical layers to ACE's 8 layers (see Table S2) with mass-weighted averages.

ERA5 dataset

We use the ERA5 reanalysis dataset spanning 1940–2022^{17,62}. Our version of the dataset—at 1° horizontal resolution and with 8 terrain-following vertical

layers—is derived from the native dataset on 137 model layers and stored in terms of spherical harmonic coefficients or on a reduced Gaussian grid, depending on the variable. It was computed from the version of ERA5 hosted by Google Research (<https://github.com/google-research/arco-era5>; ref. 63) with some auxiliary surface flux variables that were missing in the Google dataset obtained from the NCAR Research Data Archive⁶⁴. Routines from the MetView package⁶⁵ were used for the regridding. To the extent possible, data was regridded and vertically coarsened to match the SHIELD dataset's horizontal and vertical coordinate (see Table S2). Unlike the SHIELD dataset, no spherical harmonic round trip was performed on the ERA5 data because it did not show evidence of regridding artifacts in the high latitudes. Sensitivity to the vertical resolution was briefly explored: models trained on the ERA5 dataset with twice the vertical resolution—splitting the existing 8 layers into two each—were also stable with accurate time-mean climate but did not demonstrate notably improved performance (not shown).

Architecture

The Spherical Fourier Neural Operator (SFNO) architecture is used⁶⁶. This is a neural operator type architecture well suited to data on the sphere. This is the same architecture used in ref. 21. The only difference in configuration of the SFNO from version 1 of ACE is that the embedding dimension is increased from 256 to 384 for ACE2. In addition, a corrector imposing physical constraints is included as part of the model architecture, as described in the next section.

Hard physical constraints

In our previous work, we found global mean surface pressure drifted unrealistically (c.f. Fig. 9 of ref. 21). And while the model very nearly obeyed the column-wise conservation of moisture without an explicit penalty or constraint, there were still small violations of this budget and the global mean moisture budget was violated by up to 0.1 mm/day at individual time steps (c.f. Fig. 11 of ref. 21). Here we describe how we enforce hard physical constraints to eliminate these budget violations. The following equations define the budgets which we desire to impose. First, conservation of global dry air mass:

$$\langle p_s^{\text{dry}}(t + \Delta t) \rangle = \langle p_s^{\text{dry}}(t) \rangle \quad (1)$$

where $p_s^{\text{dry}}(t) = p_s(t) - gTWP(t)$ is the surface pressure due to dry air, $TWP(t) = \frac{1}{g} \int_0^p q(t, p) dp$ is the total water path, Δt is the forward time step of the machine learning model and angled brackets $\langle \rangle$ represent the area-weighted global average. Next, the conservation of column-integrated moisture:

$$\frac{TWP(t + \Delta t) - TWP(t)}{\Delta t} = E(t) - P(t) + \left. \frac{\partial TWP}{\partial t} \right|_{\text{adv}}(t) \quad (2)$$

where $E(t)$ is the evaporation rate, computed as $LHF(t)/L_v$, P is the precipitation rate and $\left. \frac{\partial TWP}{\partial t} \right|_{\text{adv}}$ is the tendency of total water path due to advection, which is directly predicted by the machine learning model (see also Table S1). Note that all of the terms on the right hand side of Eq. (2) represent time averages between t and $t + \Delta t$. Finally, we have the constraints on global moisture:

$$\left\langle \left. \frac{\partial TWP}{\partial t} \right|_{\text{adv}}(t) \right\rangle = 0 \quad (3)$$

and by implication

$$\left\langle \frac{TWP(t + \Delta t) - TWP(t)}{\Delta t} \right\rangle = \langle E(t) - P(t) \rangle. \quad (4)$$

We enforce these physical constraints on the model by including a physical corrector module within the optimized model. This module applies the following corrections to ensure the constraints are satisfied:

- (1) Moisture, precipitation rate, and radiative fluxes are all made to be positive by setting any negative values to zero.
- (2) A globally-constant surface pressure adjustment ensures total dry air mass is conserved:

$$p'_s(t) = p_s(t) - \langle p_s^{dry}(t) - p_s^{dry}(t-1) \rangle \quad (5)$$

- (3) Precipitation rate is multiplied by a constant to conserve global mean moisture:

$$P'(t) = \frac{P(t)}{\langle P(t) \rangle} \left(E(t) - \frac{TW P(t) - TW P(t-1)}{\Delta t} \right), \quad (6)$$

where $P'(t)$ is the corrected precipitation rate at time t , $P(t)$ is the precipitation prior to this correction, and $E = LHF(t)/L_v$ is the evaporation rate.

- (4) Exact conservation of column moisture is attained by deriving advective flux as residual from the adjusted TWP tendency, E and P :

$$\left. \frac{\partial TW P}{\partial t} \right|_{adv} = \frac{TW P(t) - TW P(t-1)}{\Delta t} - (E(t) - P'(t)), \quad (7)$$

where $\left. \frac{\partial TW P}{\partial t} \right|_{adv}$ represents the corrected tendency of total water path due to advection.

We avoid introducing model bias through these corrections by applying them before computing the loss. For this reason, these constraints can be considered to be part of the model architecture. The order of these adjustments is such that later corrections will not invalidate earlier corrections. These corrections are applied, by necessity, to the data in physical units instead of in normalized units.

Data normalization

For the inputs and outputs of the SFNO module, data is normalized using standard scaling. Means and standard deviations are computed over latitude, longitude and time without any area weighting. For normalization before the loss function is computed, prognostic variables are scaled to harmonize their typical difference between time steps, i.e., we use “residual” scaling (see Appendix H of ref. 21). Specifically, for a field $a(t, \phi, \lambda)$ which depends on time, latitude and longitude, the standard deviation of $a(t + \Delta t, \phi, \lambda) - a(t, \phi, \lambda)$ over time and space is used for normalization. Diagnostic variables are normalized for the loss function using standard scaling.

Table 2 | Custom weights applied to variables when computing loss function

Name	Weight
$T_0, T_1, u_0, v_0, q_0, q_2, q_{2m}, P, \left. \frac{\partial TW P}{\partial t} \right _{adv}$	0.5
q_1	0.25
$DLWRF_{sfc}, USWRF_{sfc}, DSWRF_{sfc}, USWRF_{toa}$	2
$ULWRF_{sfc}, T_{850}$	5
Z_{500}	10

Output variables which are not listed here are given a weight of 1. Variables are defined in Table S1.

For the ERA5 dataset, normalization statistics were computed over the period 1990–2020 for which this reanalysis is most reliable. For the SHIELD dataset, they were computed over 1940–2021.

Loss function

The loss function is the mean squared error over all outputs. Prognostic outputs are normalized using residual scaling as described in previous section while diagnostic outputs are normalized using standard full field scaling. The loss is summed over two autoregressive forward 6-h steps. In addition, some variables are given an additional weighting (Table 2). Variables which were downweighted are ones which showed signs of overfitting (that is, increasing 6-h RMSE on validation data late in training) without the downweighting. Variables which are upweighted are diagnostic variables, which would otherwise contribute relatively little (<0.5%) to the loss function that is averaged across 50 outputs.

Checkpoint selection based on climate skill

Since the loss function used here is based on 12-h forecast skill over two 6-hourly autoregressive steps, it is not guaranteed that a lower loss will lead to small long-term (e.g., 10-year averaged) climate biases. Since our priority in this work is accurate representation of climate statistics, we therefore define a selection criteria to choose a checkpoint with the smallest time-averaged biases. The criteria is the channel-mean global RMSE of time-means. Specifically:

$$\alpha = \frac{1}{C} \sum_{c=1}^C \sqrt{\sum_{\phi, \lambda} w_{\phi, \lambda} (\bar{y}_c(t, \phi, \lambda) - \hat{y}_c(t, \phi, \lambda))^2} \quad (8)$$

where c is an index for output channel (i.e., the prognostic and diagnostic variables), w_{ϕ} is an averaging weight proportional to area of grid cell centered at ϕ, λ . The $y_c(t, \phi, \lambda)$ is the normalized true value at a particular time and location, while \hat{y}_c is the normalized model prediction for the corresponding time, from a simulation initialized at some previous time. The overbar $\bar{\cdot}$ is a time- and ensemble-average.

In practice, α is computed once per epoch during training from an ensemble of eight 5-year long simulations, initialized at evenly spaced intervals across 1996, the start of the validation period (Table 1). In addition to choosing a best checkpoint from within a training run, we perform an ensemble of four training runs which differ only in the initialization of model parameters. For each training run, we choose a checkpoint based on minimizing α across epochs. After these training runs were completed, we found that doing inference runs over a wider span of forcing data led to a better estimate of the climate skill of a given model. Therefore to choose a checkpoint across the four training runs, we performed twelve 5-year inference runs, initialized once every 5 years starting on 1 January 1940, spanning the training and validation periods, but not overlapping with the held out test period. Additionally, for this comparison we downweighted the contribution of q_0 to the calculation of α by a factor of 10, since our poor skill in predicting the time-mean of this variable otherwise dominated α . Then the checkpoint across the four random seeds was chosen according to this new criteria. Section S1.5 shows the variability of α through training and across the four random seeds.

Table 3 | Speed and energy cost of inference with ACE2 and the physics-based SHIELD model

Model	Simulated years per day	Energy cost per simulated year [Wh]	Hardware
ACE2	1500	11.2	1 NVIDIA H100-80GB-HBM
C24 SHIELD	22.1	300	54 cores on 1 AMD EPYC 7H12
C96 SHIELD	11.4	8250	864 cores on 14 AMD EPYC 7H12

ACE2 and C96 SHIELD both have about 1° horizontal resolution while C24 SHIELD has about 4° resolution but is still more than an order of magnitude more energy-intensive than ACE2.

Evaluation metrics

To evaluate time-mean climate skill, we compute the global RMSE of the time-mean for an individual variable y as:

$$\sqrt{\sum_{\phi,\lambda} w_{\phi,\lambda} (\overline{y(t, \phi, \lambda)} - \hat{y}(t, \phi, \lambda))^2} \quad (9)$$

where $w_{\phi,\lambda}$ is an area weight that sums to 1 and \hat{y} is the predicted value and the overline represents a time average. For global- and annual-mean series of a given output variable, we also compute an R^2 of the predicted series against a reference series of that variable:

$$R^2 = 1 - \frac{SS_{error}}{SS_{reference}} \quad (10)$$

where:

$$SS_{error} = \sum_{i_{year}=1}^{n_{years}} (\hat{y}_{i_{year}} - y_{i_{year}})^2 \quad (11)$$

and

$$SS_{reference} = \sum_{i_{year}=1}^{n_{years}} (y_{i_{year}} - \bar{y})^2. \quad (12)$$

Here \hat{y} and y are predicted and reference variable values, and \bar{y} is the average over the time period. Thus R^2 reflects the model's combined ability to capture long-term means and trends as well as shorter-term inter-annual variability.

To characterize the atmospheric response to El Niño-Southern Oscillation (ENSO) we compute a regression coefficient of variables against the historical Niño 3.4 index²⁸ as computed from the CMIP6 AMIP SST dataset^{18,57,58}. The coefficient is β_1 of a linear regression:

$$\hat{y} = \beta_1 I_{Niño34} + \beta_0 \quad (13)$$

where $I_{Niño34}$ is the 3-monthly centered running mean of SSTs in the Niño 3.4 region, after being nearest-neighbor interpolated to the 6-hourly time frequency of data. This produces a map of the response of a particular variable to seasonally-varying ENSO states. We compare the predicted response against a reference dataset response by computing the global area-weighted RMS difference between the response maps. This also allows for computing the variability of the SHIELD reference dataset's atmospheric response to ENSO, as the difference between the response maps of its two initial conditions.

Computational cost

Training duration for each model is approximately 4.5 days on eight NVIDIA H100-80GB-HBM3 GPUs. For each dataset, four models were trained with the same hyperparameters and differing only in parameter initialization (see "Checkpoint selection based on climate skill" section above) quadrupling the overall cost. The cost of doing inference with ACE2 and the reference SHIELD model is shown in Table 3. Comparing ACE2 and C96 SHIELD, which have the same horizontal resolution, ACE2 is about 100 times faster and 700 times less energy intensive. Even compared to C24 SHIELD, which has four times lower horizontal resolution, ACE2 uses about 25 times less energy and is about 50 times faster.

Data availability

The ERA5 dataset is available from the Copernicus Climate Data Store (<https://cds.climate.copernicus.eu/>). The processed version of the dataset used to train ACE2-ERA5 is available on a public requester-pays Google

Cloud Storage bucket at <gs://ai2cm-public-requester-pays/2024-11-13-ai2-climate-emulator-v2-amip/data/era5-1deg-1940-2022.zarr> (about 1.5TiB). Similarly, the SHIELD dataset used to train ACE2-SHIELD is available at <gs://ai2cm-public-requester-pays/2024-11-13-ai2-climate-emulator-v2-amip/data/c96-1deg-shield> (about 3 TiB).

Code availability

The code used for data processing, model training, inference and evaluation is available at <https://github.com/ai2cm/ace> (ref. 67). The trained ACE2-ERA5 model checkpoint is available at <https://doi.org/10.57967/hf/5377> (ref. 68). The scripts used for submitting experiments and generating figures are available at <https://github.com/ai2cm/ace2-paper> (ref. 69).

Received: 4 December 2024; Accepted: 13 May 2025;

Published online: 29 May 2025

References

- Kay, J. E. et al. The community earth system model (CESM) large ensemble project: a community resource for studying climate change in the presence of internal climate variability. *Bull. Am. Meteorol. Soc.* **96**, 1333–1349 (2015).
- Mahesh, A. et al. Huge ensembles part I: design of ensemble weather forecasts using spherical Fourier neural operators. <https://arxiv.org/abs/2408.03100> (2024).
- Milinski, S., Maher, N. & Olonscheck, D. How large does a large ensemble need to be? *Earth Syst. Dyn.* **11**, 885–901 (2020).
- Claussen, M. et al. Earth system models of intermediate complexity: closing the gap in the spectrum of climate system models. *Clim. Dyn.* **18**, 579–586 (2002).
- Watson-Parris, D. et al. Climatebench v1.0: a benchmark for data-driven climate projections. *J. Adv. Model. Earth Syst.* **14**, <https://doi.org/10.1029/2021MS002954> (2022).
- Brajard, J., Carrassi, A., Bocquet, M. & Bertino, L. Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: a case study with the Lorenz 96 model. *J. Comput. Sci.* **44**, 101171 (2020).
- Hatfield, S. et al. Building tangent-linear and adjoint models for data assimilation with neural networks. *J. Adv. Model. Earth Syst.* **13**, <https://doi.org/10.1029/2021MS002521> (2021).
- Perkins, W. A. & Hakim, G. J. Coupled atmosphere-ocean reconstruction of the last millennium using online data assimilation. *Paleoceanogr. Paleoclimatol.* **36**, <https://doi.org/10.1029/2020PA003959> (2021).
- Bi, K. et al. Accurate medium-range global weather forecasting with 3d neural networks. *Nature* **619**, 533–538 (2023).
- Lam, R. et al. Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).
- Price, I. et al. Gencast: diffusion-based ensemble forecasting for medium-range weather. <https://arxiv.org/abs/2312.15796> (2023).
- Chen, L. et al. A machine learning model that outperforms conventional global subseasonal forecast models. *Nat. Commun.* **15**, <https://doi.org/10.1038/s41467-024-50714-1> (2024).
- Kochkov, D. et al. Neural general circulation models for weather and climate. *Nature* **632**, 1060–1066 (2024).
- Kochkov, D. et al. Machine learning-accelerated computational fluid dynamics. *Proc. Natl Acad. Sci. USA* **118**, e2101784118 (2021).
- Beucler, T. et al. Climate-invariant machine learning. *Sci. Adv.* **10**, eadj7250 (2024).
- Clark, S. K. et al. Correcting a 200 km resolution climate model in multiple climates by machine learning from 25 km resolution simulations. *J. Adv. Model. Earth Syst.* **14**, <https://doi.org/10.1029/2022MS003219> (2022).
- Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).

18. Eyring, V. et al. Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.* **9**, 1937–1958 (2016).
19. Harris, L. et al. GFDL Shield: a unified system for weather-to-seasonal prediction. *J. Adv. Model. Earth Syst.* **12**, <https://doi.org/10.1029/2020MS002223> (2020).
20. Clark, S. K. et al. Ace2-som: coupling an ml atmospheric emulator to a slab ocean and learning the sensitivity of climate to changed CO₂. <https://arxiv.org/abs/2412.04418> (2024).
21. Watt-Meyer, O. et al. ACE: a fast, skillful learned global atmospheric model for climate prediction. <https://arxiv.org/abs/2310.02074> (2023).
22. Duncan, J. P. C. et al. Application of the ai2 climate emulator to e3smv2's global atmosphere model, with a focus on precipitation fidelity. *J. Geophys. Res. Mach. Learn. Comput.* **1**, <https://doi.org/10.1029/2024JH000136> (2024).
23. Karlbauer, M. et al. Advancing parsimonious deep learning weather prediction using the healpix mesh. *J. Adv. Model. Earth Syst.* **16**, <https://doi.org/10.1029/2023MS004021> (2024).
24. Guan, H., Arcomano, T., Chattopadhyay, A. & Maulik, R. LUCIE: a lightweight uncoupled climate emulator with long-term stability and physical consistency for O(1000)-member ensembles. <https://arxiv.org/abs/2405.16297> (2024).
25. Cresswell-Clay, N. et al. A deep learning earth system model for stable and efficient simulation of the current climate. <https://arxiv.org/abs/2409.16247> (2024).
26. Rühling Cachay, S., Henn, B., Watt-Meyer, O., Bretherton, C. S. & Yu, R. Probabilistic emulation of a global climate model with spherical dyffusion. <https://arxiv.org/abs/2406.14798> (2024).
27. Watt-Meyer, O. et al. Neural network parameterization of subgrid-scale physics from a realistic geography global storm-resolving simulation. *J. Adv. Model. Earth Syst.* **16**, <https://doi.org/10.1029/2023MS003668> (2024).
28. Trenberth, K. E. The definition of El Niño. *Bull. Am. Meteorol. Soc.* **78**, 2771–2777 (1997).
29. Elsner, J. B., Kossin, J. P. & Jagger, T. H. The increasing intensity of the strongest tropical cyclones. *Nature* **455**, 92–95 (2008).
30. Bhatia, K. T. et al. Recent increases in tropical cyclone intensification rates. *Nat. Commun.* **10**, <https://doi.org/10.1038/s41467-019-08471-z> (2019).
31. Vecchi, G. A., Landsea, C., Zhang, W., Villarini, G. & Knutson, T. Changes in atlantic major hurricane frequency since the late-19th century. *Nat. Commun.* **12**, <https://doi.org/10.1038/s41467-021-24268-5> (2021).
32. Hodges, K., Cobb, A. & Vidale, P. L. How well are tropical cyclones represented in reanalysis datasets? *J. Clim.* **30**, 5243–5264 (2017).
33. Ullrich, P. A. et al. TempestExtremes v2.1: a community framework for feature detection, tracking, and analysis in large datasets. *Geosci. Model Dev.* **14**, 5023–5048 (2021).
34. Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J. & Neumann, C. J. The international best track archive for climate stewardship (IBTrACS): unifying tropical cyclone data. *Bull. Am. Meteorol. Soc.* **91**, 363–376 (2010).
35. Kenneth, R., Howard, J., James, P., Michael, C. & Carl, J. International best track archive for climate stewardship (IBTrACS) project, version 4. <https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C01552> (2019).
36. Price, J. F. Upper ocean response to a hurricane. *J. Phys. Oceanogr.* **11**, 153–175 (1981).
37. Golaz, J. et al. The DOE E3SM model version 2: overview of the physical model and initial model evaluation. *J. Adv. Model. Earth Syst.* **14**, <https://doi.org/10.1029/2022MS003156> (2022).
38. Ahn, M. et al. MJO propagation across the maritime continent: are CMIP6 models better than CMIP5 models? *Geophys. Res. Lett.* **47**, <https://doi.org/10.1029/2020GL087250> (2020).
39. Hagos, S., Leung, L. R. & Dudhia, J. Thermodynamics of the Madden-Julian oscillation in a regional model with constrained moisture. *J. Atmos. Sci.* **68**, 1974–1989 (2011).
40. Wheeler, M. & Kiladis, G. N. Convectively coupled equatorial waves: analysis of clouds and temperature in the wavenumber-frequency domain. *J. Atmos. Sci.* **56**, 374–399 (1999).
41. Zhang, C. Madden-Julian oscillation. *Rev. Geophys.* **43**, <https://doi.org/10.1029/2004RG000158> (2005).
42. Waliser, D. et al. MJO simulation diagnostics. *J. Clim.* **22**, 3006–3030 (2009).
43. Waliser, D. E. et al. Agcm simulations of intraseasonal variability associated with the asian summer monsoon. *Clim. Dyn.* **21**, 423–446 (2003).
44. Kim, D. et al. Application of MJO simulation diagnostics to climate models. *J. Clim.* **22**, 6413–6436 (2009).
45. Weyn, J. A., Durran, D. R. & Caruana, R. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *J. Adv. Model. Earth Syst.* **12**, <https://doi.org/10.1029/2020MS002109> (2020).
46. Chen, L. et al. Fuxi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Clim. Atmos. Sci.* **6**, <https://doi.org/10.1038/s41612-023-00512-1> (2023).
47. Baldwin, M. P. & Dunkerton, T. J. Stratospheric harbingers of anomalous weather regimes. *Science* **294**, 581–584 (2001).
48. Solomon, S. Stratospheric ozone depletion: a review of concepts and history. *Rev. Geophys.* **37**, 275–316 (1999).
49. Anstey, J. A. et al. Impacts, processes and projections of the quasi-biennial oscillation. *Nat. Rev. Earth Environ.* **3**, 588–603 (2022).
50. Rasp, S. et al. Weatherbench 2: a benchmark for the next generation of data-driven global weather models. *J. Adv. Model. Earth Syst.* **16**, <https://doi.org/10.1029/2023MS004019> (2024).
51. Manabe, S. & Wetherald, R. T. Thermal equilibrium of the atmosphere with a given distribution of relative humidity. *J. Atmos. Sci.* **24**, 241–259 (1967).
52. Screen, J. A., Deser, C. & Simmonds, I. Local and remote controls on observed arctic warming. *Geophys. Res. Lett.* **39**, <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012GL051598> (2012).
53. Urraca, R. et al. Evaluation of global horizontal irradiance estimates from era5 and cosmo-rea6 reanalyses using ground and satellite-based data. *Sol. Energy* **164**, 339–354 (2018).
54. Zhou, L. et al. Improving global weather prediction in GFDL SHIELD through an upgraded GFDL cloud microphysics scheme. *J. Adv. Model. Earth Syst.* **14**, e2021MS002971 (2022).
55. Cheng, K.-Y. et al. Impact of warmer sea surface temperature on the global pattern of intense convection: insights from a global storm resolving model. *Geophys. Res. Lett.* **49**, e2022GL099796 (2022).
56. Deser, C., Phillips, A., Bourdette, V. & Teng, H. Uncertainty in climate change projections: the role of internal variability. *Clim. Dyn.* **38**, 527–546 (2012).
57. Taylor, K. E., Williamson, D. & Francis, Z. The Sea Surface Temperature and Sea-ice Concentration Boundary Conditions for AMIP II Simulations. PCMDI Rep. 60, *Lawrence Livermore National Laboratory* (2000).
58. Durack, P. J., Taylor, K. E., Po-Chedley, S. & Doutriaux, C. AMIPBCs — AMIP dataset prepared for input4MIPS. <https://github.com/PCMDI/amipbcs/releases/tag/1.1.8> (2022).
59. Meinshausen, M. et al. Historical greenhouse gas concentrations for climate modelling (CMIP6). *Geosci. Model Dev.* **10**, 2057–2116 (2017).
60. Conway, T. J. et al. Evidence for interannual variability of the carbon cycle from the National Oceanic and Atmospheric Administration/Climate Monitoring and Diagnostics Laboratory Global Air Sampling Network. *J. Geophys. Res. Atmos.* **99**, 22831–22855 (1994).
61. NOAA-GFDL. NOAA-GFDL/FRE-NCtools. NOAA—Geophysical Fluid Dynamics Laboratory. <https://github.com/NOAA-GFDL/FRE-NCtools> (2024).

62. Soci, C. et al. The ERA5 global reanalysis from 1940 to 2022. *Q. J. R. Meteorol. Soc.* **150**, 4014–4048 (2024).
63. Carver, R. W. & Merosse, A. ARCO-ERA5: an analysis-ready cloud-optimized reanalysis dataset. In *22nd Conf. on AI for Env. Science*, Denver, CO, *Amer. Meteor. Soc.* <https://ams.confex.com/ams/103ANNUAL/meetingapp.cgi/Paper/415842> (2023).
64. European Centre for Medium-Range Weather Forecasts. ERA5 reanalysis (0.25 degree latitude-longitude grid) <https://rda.ucar.edu/datasets/dsd633000/> (2019).
65. Russell, I. & Kertész, S. Metview. <https://www.ecmwf.int/node/18136> (2017).
66. Bonev, B. et al. Spherical fourier neural operators: learning stable dynamics on the sphere. In *Proceedings of the 40th International Conference on Machine Learning (ICML)* (2023).
67. McGibbon, J. et al. ACE software v2024.12.0 <https://zenodo.org/doi/10.5281/zenodo.14503970> (2024).
68. Watt-Meyer, O. et al. ACE2-ERA5 model checkpoint. <https://huggingface.co/allenai/ACE2-ERA5> (2025).
69. Henn, B., McGibbon, J. & Watt-Meyer, O. ACE2 paper manuscript figures v2025.05.01. <https://zenodo.org/doi/10.5281/zenodo.15368148> (2025).

Acknowledgements

We acknowledge NOAA's Geophysical Fluid Dynamics Laboratory for providing the computing resources used to perform the reference SHIELD simulations. This research used resources of NERSC, a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, using NERSC award BER-ERCAP0026743. We acknowledge ECMWF for generating and providing the ERA5 dataset.

Author contributions

O.W., B.H., J.M. and C.B. conceived of the study and designed the experiments. O.W., B.H., J.M., S.C. and W.A.P. generated and processed the datasets. O.W., B.H., S.C., A.K. and J.M. performed the experiments, prepared the figures and wrote the manuscript. O.W., B.H., J.M., S.C., A.K.,

W.A.P. and E.W. contributed to software to enable the experiments and analysis. All authors reviewed the manuscript and provided feedback.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41612-025-01090-0>.

Correspondence and requests for materials should be addressed to Oliver Watt-Meyer.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025