npj | climate and atmospheric science

Article

Published in partnership with CECCR at King Abdulaziz University

https://doi.org/10.1038/s41612-026-01585-4

Generative assimilation and prediction for weather and climate

Check for updates

Shangshang Yang[1,2,3,4,16], Congyi Nai[1,16], Xinyan Liu[1,5], Weidong Li[1,15], Jie Chao[1], Jingnan Wang[1,6], Leyi Wang[7,8], Lu Li[9], Xichen Li[10,11], Xi Chen[1], Bo Lu[12,13], Xin Li[14], Ziniu Xiao[1], Niklas Boers[3,4] ✉, Huiling Yuan[2] ✉ & Baoxiang Pan[1] ✉

Machine learning models have shown great success in predicting weather up to two weeks ahead, outperforming process-based benchmarks[1–5]. However, most existing approaches do not fully integrate data assimilation into the prediction workflow[6,7], and often suffer from error accumulation in autoregressive computation, limiting their applicability to predictions beyond the medium range[8] and climate projections[9]. We introduce Generative Assimilation and Prediction (GAP), a unified deep generative framework that connects data assimilation, weather forecasting, and climate simulation through a shared probabilistic representation of atmospheric states. GAP learns the climatological distribution from reanalysis data, and adapts it to different tasks by incorporating observational, predictive, and forcing constraints. Evaluated across a range of weather-climate tasks, GAP produces probabilistic weather forecasts competitive with leading forecasting systems, using its own assimilated initial states derived from 5% synthetic perfect observations, with particular skill in capturing extreme events such as atmospheric rivers, heatwaves and the Madden-Julian Oscillation. At seasonal timescales, GAP shows promising predictive skill when constrained by oceanic boundary forcing. Moreover, GAP supports stable millennial-scale climate simulations that reproduce atmospheric variability from daily weather fluctuations to decadal oscillations. These results suggest that generative modeling of the climatological distribution offers a viable foundation for seamless weather and climate prediction, thereby opening a path toward integrated, observation-to-climate workflows.

Reliable modeling of Earth's chaotic atmosphere is fundamental to predicting both day-to-day weather and long-term climate change. Recent decades have seen steady progress in our understanding and simulation of atmospheric dynamics[10]. These improvements have enhanced forecasting through two key processes: data assimilation (Fig. 1A left), which combines observations with models to infer current atmospheric states[11]; and prediction (Fig. 1A right), which projects the initial state estimates forward in time, until chaos inevitably blurs our view of the future[12]. At the core of both processes lies a fundamental challenge — how to accurately represent and evolve the probabilities of possible atmospheric states, given sparse observations, imperfect prediction models, and inconspicuous natural and anthropogenic forcings[13]. Successfully addressing this challenge would bridge a long-standing gap between weather forecasting and climate projection, connecting our

[1]Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China. [2]School of Atmospheric Sciences and Key Laboratory of Mesoscale Severe Weather/Ministry of Education, Nanjing University, Nanjing, China. [3]Earth System Modelling, School of Engineering and Design, Technical University of Munich, Munich, Germany. [4]Complexity Science, Potsdam Institute for Climate Impact Research, Potsdam, Germany. [5]Chinese Academy of Meteorological Sciences, Beijing, China. [6]College of Meteorology and Oceanography, National University of Defense Technology, Changsha, China. [7]Hubei Center for Applied Mathematics, Wuhan University, Wuhan, China. [8]Academy of Advanced Interdisciplinary Studies, Wuhan University, Wuhan, China. [9]School of Atmospheric Sciences, Sun Yat-Sen University, Zhuhai, China. [10]Institute of Ocean Research, School of Physics, Peking University, Beijing, China. [11]Department of Atmospheric and Oceanic Sciences, School of Physics, Peking University, Beijing, China. [12]China Meteorological Administration Key Laboratory for Climate Prediction Studies, National Climate Center, Beijing, China. [13]Xiong'an Institute of Meteorological Artificial Intelligence, Xiong'an, China. [14]Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing, China. [15]Present address: Department of the Geophysical Sciences, University of Chicago, Chicago, USA. [16]These authors contributed equally: Shangshang Yang, Congyi Nai. ✉e-mail: n.boers@tum.de; yuanhl@nju.edu.cn; panbaoxiang@lasg.iap.ac.cn

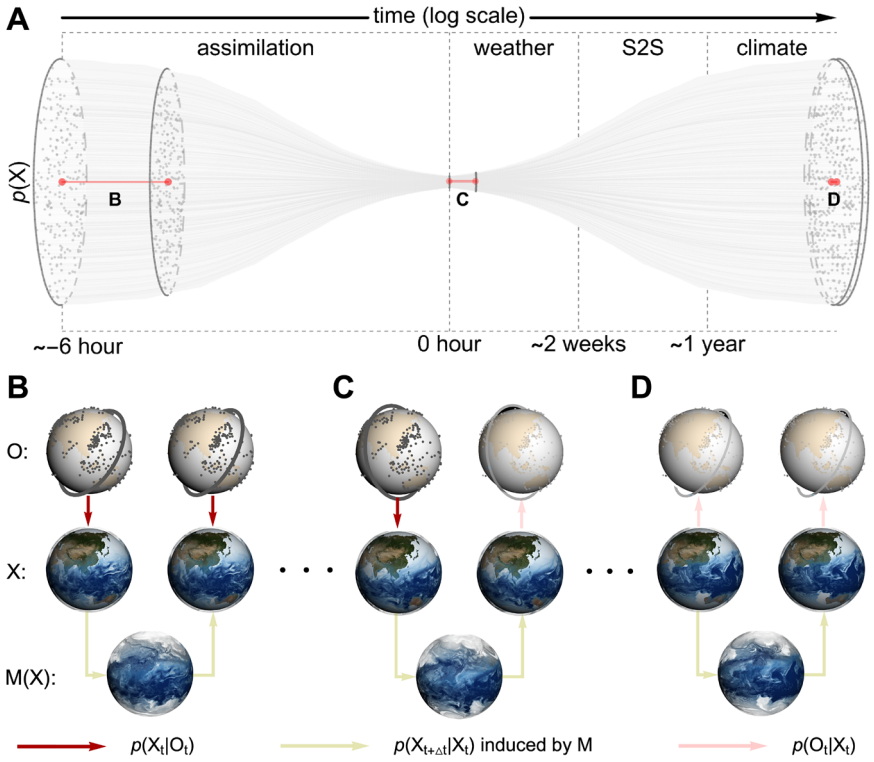


Fig. 1 | Schematic of the Generative Assimilation and Prediction (GAP) framework. A Evolution of atmospheric state probability distributions, illustrating the Probabilistic Assimilation and Prediction (PAP) concept. A logarithmic timeline is adopted to highlight the duality between assimilation and prediction that we aim to unify[32]. During assimilation (left), observations progressively constrain the climatological distribution, narrowing possible states. During prediction (right), uncertainty evolves through distinct phases: weather forecasting (up to around two weeks) shows growing uncertainty until deterministic predictions lose utility; seasonal prediction (up to around 1–2 years) maintains predictability below climatological levels through low-frequency climate process constraints; and climate prediction (beyond 1–2 years) shows complete dissipation of initial condition information while external forcings modify the underlying climatological distribution. **B**–**D** Practical implementation of GAP. **B** Assimilation phase implementation: given state estimates sampled from $p(\mathbf{X}_t|\mathbf{X}_{t-\Delta t}, \mathbf{O}_t)$ --- the probability distribution of the current state $\mathbf{X}_t$ conditioned on the previous state $\mathbf{X}_{t-\Delta t}$ (or climatological distribution if $\mathbf{X}_{t-\Delta t}$ were unavailable) and current observations $\mathbf{O}_t$ --- we use a forecasting model **M** to generate preliminary predictions, which together with observations $\mathbf{O}_{t+\Delta t}$ guide subsequent probabilistic state estimation for $p(\mathbf{X}_{t+\Delta t}|\mathbf{X}_t, \mathbf{O}_{t+\Delta t})$. **C** Transition point marking the end of assimilation and beginning of prediction, where forecasts proceed without observational input while yielding predicted observations via $p(\mathbf{O}|\mathbf{X})$. **D** Prediction phase implementation: preliminary model predictions are combined with climatological priors to sample possible states across weather-to-climate timescales, maintaining physical consistency throughout.

understanding of individual atmospheric processes[14] to the climate system's overall behavior[15]. Despite steady progress, accurately tracking these probability distributions remains difficult due to model biases, the curse of dimensionality, and computational constraints.

Machine learning (ML) has emerged as a transformative avenue for weather forecasting[1–5], achieving accurate predictions up to 2 weeks ahead with unprecedented computational efficiency. Yet, most existing approaches have primarily emphasized prediction accuracy[16–19], with comparatively less attention to other essential components of the prediction workflow. In particular, these models typically depend on initial conditions provided by separate data assimilation systems and lack mechanisms to incorporate new observations as they become available. Meanwhile, they struggle to maintain stability over the longer integration horizons required for seasonal prediction and climate projection. Efforts to address these gaps are progressing on multiple fronts. ML-based data assimilation methods have gained considerable momentum[20–24], though existing implementations often suffer from inadequate uncertainty quantification[12], or face challenges when scaling to high-dimensional operational settings[22,24]. Extensions toward longer temporal ranges have created new opportunities for seasonal prediction[8] and climate simulation[9,25], yet improved stability often comes at the cost of prediction skill[9], underestimation of extremes[8,26], and many models fail to capture the intrinsic stochasticity essential to climate dynamics[25]. Taken together, these developments reveal a broader structural limitation: current ML methods often address individual components of the weather-climate modeling pipeline without recognizing that they are manifestations of a single underlying challenge: how to represent and evolve the probability distribution of atmospheric states across all timescales.

The fundamental pursuit of climate science extends beyond efficient prediction or emulation — it seeks a comprehensive understanding of how the climate system evolves as a chaotic, high-dimensional, multi-scale system under various forcings[13,27,28]. Achieving such an understanding requires an inherently probabilistic perspective, as neither observation nor model is comprehensive or errorless[29]. Consequently, the atmospheric state cannot be known with certainty, but only in terms of a distribution over possible states[30,31]. Yet despite this complexity and uncertainty, the atmosphere exhibits structured patterns across scales, giving rise to a characteristic climatological distribution: the set of all possible states weighted by their probability of occurrence. This distribution encodes both natural variability and the system's response to external forcings.

From this perspective, assimilation and prediction can be viewed as two complementary operations on the distribution (Fig. 1A[32]). Observations constrain it toward states consistent with the available observations, while prediction advances it forward in time. As chaotic dynamics progressively erase memory of the initial state, the distribution relaxes back toward climatology across weather, subseasonal-to-seasonal, and climate timescales (Fig. 1A). Realizing such Probabilistic Assimilation and Prediction (PAP, see "Methods") in practice has proven difficult. Conventional methods rely on Gaussian assumptions and linear approximations that poorly represent the high-dimensional, nonlinear structure of the atmosphere[33,34]. Stochastic parameterizations[35] and perturbation-based ensemble approaches[36] offer partial remedies, yet they remain tailored to specific timescales. Recent advances in deep generative models[37–39] — applied separately to data assimilation[22,40–42], ensemble forecasting[4], and climate simulation[43,44] — suggest a more direct path forward. Yet these efforts remain narrowly

focused on individual tasks, rather than providing a unified representation of the climatological distribution across all scales[45].

Here we present GAP, a framework that implements this unified probabilistic view using deep generative models. GAP learns the climatological distribution of atmospheric states from reanalysis data, and adapts it to different tasks — assimilation, forecasting, and climate simulation — by incorporating observational, predictive, and forcing constraints at inference time, without requiring separate models or training objectives for each task.

Results

Generative assimilation and prediction

We formalize the unified probabilistic view described above as PAP (Fig. 1A): the ideal Bayesian target is the conditional distribution of atmospheric states given observational, predictive, and forcing constraints; assimilation progressively narrows this distribution as observations arrive, while prediction allows it to evolve back toward climatology as chaos erases initial-state information. GAP provides a practical, approximate realization of PAP using diffusion-based inference, chaining two components (Fig. 1B–D):

1. A *generative model* — a probabilistic diffusion model [37–39] trained on atmospheric reanalysis (Supplementary Section A) to approximate the climatological distribution of atmospheric states (Supplementary Section B.1). The model learns this distribution by iteratively reversing a noise-adding process, and at inference time flexibly incorporates observational, predictive, and external forcing constraints via approximate Bayesian computation[39,46–50].
2. A *forecasting model* that propagates the state estimate forward in time through both the assimilation and prediction stages (Supplementary Section B.2). This component can be either numerical or data-driven; we demonstrate GAP with the latter for implementation convenience, though either is viable in principle.

In practice, these two components form a coherent cycle. At each assimilation step (Fig. 1B), observations and preliminary predictions constrain the generative model using modified diffusion posterior sampling methods [46,47,49,50], (see "Methods"). These constrained estimates are then evolved forward by the forecasting model, creating preliminary predictions for the next time step. Rather than directly using these predictions, GAP applies calibrated noise and reverses the diffusion process[43,49], effectively introducing stochasticity that adheres to the learned phase space structure, while simultaneously correcting systematic biases in the forecasting model[51].

As we transition to prediction (Fig. 1C, D), this cycle continues without observational constraints, allowing the well-constrained initial state distribution to gradually drift toward the climatological distribution. This evolution passes through distinct phases: weather forecasting (up to 2 weeks), where initial conditions dominate; subseasonal-to-seasonal prediction (up to 1 year), where ocean-atmosphere coupling provides extended predictability; and climate projection, where external forcings modulate the underlying climatological distribution. For these longer timescales, we leverage sea surface temperatures as a simplified external forcing constraint capturing both seasonal memory and climate change signals, effectively translating oceanic influences into appropriate atmospheric responses (Supplementary Sections B.4 and B.5). This unified framework maintains probabilistic rigor and physical consistency across all prediction horizons without requiring separate training for different timescales. An ablation study demonstrating the contribution of each component of GAP is provided in Supplementary Section D.7.

We evaluate GAP's capabilities across five key aspects in weather-climate modeling: (1) its ability to reproduce climatological patterns and variability; (2) its effectiveness in data assimilation for establishing accurate initial conditions; (3) its skill in weather forecasting; (4) its potential for seasonal prediction; and (5) its capacity to generate stable long-term climate simulations and projections. Details of the evaluation metrics can be found in Supplementary Section C.

Climatology

The GAP framework rests on the premise that a generative model can faithfully approximate the climatological distribution of atmospheric states. Before evaluating on assimilation and prediction tasks, we must therefore assess whether the generative model captures this distribution with sufficient fidelity. As direct comparison of high-dimensional distributions is intractable, we follow established strategies developed for evaluating physical climate models[52]: we generate 5,000 unconditional samples and compare their statistical properties against ERA5 reanalysis[53] across the training period (1979–2017), progressing from grid-point statistics to large-scale dynamical structures to inter-variable physical consistency. To provide context for interpreting GAP's performance, we include a parallel evaluation of the Community Earth System Model version 2 (CESM2[54]) — a state-of-the-art physical climate model — using the same diagnostics (Supplementary Fig. 4).

At the grid-point level, GAP accurately reproduces the statistical properties of individual atmospheric variables: Fig. 2A shows that the spatial patterns of mean, variance, minimum, and maximum of 2-m temperature closely match those of ERA5. Stippled regions show where GAP's generated distributions are statistically indistinguishable from ERA5 at the 95% confidence level, according to a Kolmogorov–Smirnov test (Supplementary Section C.4), confirming that the generative model captures the full local probability distribution rather than just the central tendency. This fidelity extends across all variables and pressure levels, with typical deviations below 10% for most metrics (Supplementary Fig. 5; spatial patterns for additional variables in Supplementary Figs. 6–10). By comparison, CESM2 reproduces the broad spatial patterns of 2-m temperature statistics but exhibits larger systematic deviations in variance and extremes (Supplementary Fig. 4A), consistent with the known biases of physical climate models in representing regional-scale variability.

Beyond point-wise statistics, GAP reproduces the organized spatial and dynamical structures of the atmosphere. The leading empirical orthogonal function of 500 hPa geopotential height (Fig. 2B) shows that GAP captures the Northern Annular Mode — a dominant pattern governing jet stream position and storm tracks. CESM2 also captures this mode, though with reduced amplitude compared with ERA5 (Supplementary Fig. 4B). This large-scale accuracy extends to the vertical structure of the general circulation: GAP faithfully reproduces the Hadley cell (Fig. 2C), including both the tropical cell's strength and vertical extent. CESM2 represents the Hadley circulation with correct gross meridional extent but shows notable biases in vertical structure and zonal-mean wind strength (Supplementary Fig. 4C).

Across spatial scales, the spherical harmonic power spectra (Fig. 2D) confirm that GAP preserves the characteristic power-law decay from planetary to regional scales (Supplementary Figs. 11 and 12 for additional variables). CESM2 reproduces the expected spectral slope but shows excess power at intermediate wavelengths (Supplementary Fig. 4), indicating a mismatch in synoptic-scale variability that is largely absent in GAP's output.

Finally, the model maintains physically consistent relationships among variables. The vertical profiles of ageostrophic-to-geostrophic wind ratios in the extratropics (Fig. 2E) are nearly identical between GAP and ERA5, indicating that the generative model preserves fundamental atmospheric balance conditions without these being explicitly imposed. CESM2 shows reasonable large-scale balance in this diagnostic, but with systematic deviations in the upper troposphere (Supplementary Fig. 4).

This multi-scale agreement — spanning grid-point distributions, large-scale circulation patterns, and inter-variable physical relationships — demonstrates that the generative model yields a reasonable approximation of the climatological atmospheric distribution. These results position GAP as a promising foundation for the assimilation and prediction tasks that follow.

Assimilation

We now evaluate whether GAP can use the learned climatology prior to infer atmospheric states from sparse observations — the data assimilation

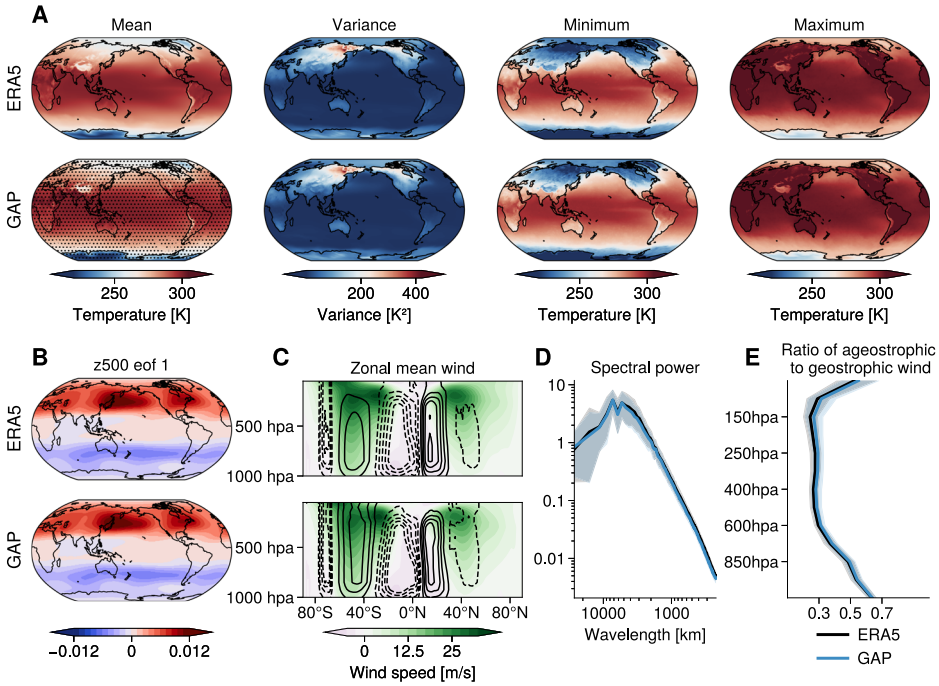


Fig. 2 | Multi-scale evaluation of GAP's climatological representation. A Grid-point scale comparison between ERA5 and GAP using 2-m temperature statistics (mean, variance, minimum, and maximum). Stippled areas indicate regions where GAP's generated distributions are statistically indistinguishable from ERA5 at the 95% confidence level according to a Kolmogorov–Smirnov test. **B** Leading empirical orthogonal function (EOF) of 500 hPa geopotential height, revealing GAP's accurate reproduction of the Northern Annular Mode pattern. **C** Hadley cell circulation represented by mass streamfunction (contours) and zonal-mean winds (colors) as a function of pressure and latitude. Solid (dashed) contours indicate positive (negative) values of the stream function, with contour intervals set to 2×10^{10} kg/s. **D** Spatial power spectral density as function of wavelength for 2-m temperature, showing the characteristic power law decay from global to local scales. **E** Vertical profiles of ageostrophic-to-geostrophic wind ratio in the extratropics (|lat|≥20°), demonstrating preservation of fundamental atmospheric balance conditions.

task[11]. As denoted earlier, GAP performs assimilation through sequential Bayesian inference: at each time step, the generative model updates its estimate of the atmospheric state by conditioning on both predictive information and incoming observations (Fig. 1B), analogous in spirit to ensemble Kalman or particle filters but using a generative model for sampling and inference (see "Methods").

We design two complementary experiments to test this capability under different levels of observational idealization. GAP-OSSE (Observing System Simulation Experiment) uses synthetic "perfect" observations drawn from ERA5 reanalysis, isolating GAP's assimilation performance from complications introduced by real-world measurement error[55]. GAP-OBS assimilates actual observations from global radiosondes and surface stations (Supplementary Section A.2), which are irregularly distributed and contain measurement uncertainties. It is worth noticing that GAP-OBS is not designed as an operational system for now. It serves purely as a preliminary proof-of-concept test of whether the framework can handle the complexities of real observational data.

A central question, echoing the probabilistic emphasis of the proposed framework, is whether GAP's ensemble representation provides reasonably calibrated uncertainty estimates. We illustrate this with a representative case at 00 UTC on 3 March 2023 (Fig. 3A–H). In both GAP-OSSE and GAP-OBS experiments, the ensemble mean fields converge to closely match the ERA5 reference state (Fig. 3A–D). More importantly, the spatial patterns of ensemble spread (Fig. 3E, F) show strong correspondence with the actual ensemble mean error (Fig. 3G, H). This alignment indicates that GAP's uncertainty estimates are informative rather than arbitrary: regions where the ensemble is more uncertain tend to be regions where the analysis is less accurate, and vice versa. Comparing the two experiments, GAP-OSSE exhibits slightly tighter spread and error patterns, as expected from its idealized observations, while GAP-OBS demonstrates robustness to real-world observational uncertainties without substantial degradation in reliability. A more formal assessment using coverage probabilities and rank histograms (Supplementary Figs. 16 and 17) confirms that the ensembles are informative and approximately calibrated, though not perfectly: the narrower prediction intervals are mildly under-covered, and the rank histograms show mild underdispersion with the truth preferentially falling below the ensemble, a deficiency that diminishes with DA cycling.

We further examine how observational information propagates spatially: Fig. 3I shows that both ensemble variance and ensemble mean error decrease near observation locations and increase with distance, following a consistent spatial decay pattern (Supplementary Fig. 14). This behavior is physically expected — observations should most strongly constrain nearby states — and its presence in GAP's output indicates that the generative model respects the spatial correlation structure of atmospheric fields rather than treating observations as isolated constraints.

A systematic evaluation across observation densities (Fig. 3J, and Supplementary Fig. 15) reveals the interplay between observational coverage and the atmosphere's chaotic tendency to amplify uncertainty. As observation density increases from 0 (climatological level) to 10^4 points (~15% global grid coverage), the analysis converges more rapidly to the reference state. At low densities, multiple assimilation cycles are needed to accumulate sufficient constraints, while dense observations can constrain the state almost immediately. This sensitivity to observation count is consistent with the theoretical expectation that denser measurements more effectively counteract chaotic error growth.

While these diagnostics assess the internal consistency of GAP's assimilation, the most stringent test of any analysis system is whether the resulting initial conditions support skillful forecasts. We turn to this evaluation next.

Weather forecast

GAP generates an ensemble forecast following a similar prediction-refine process as in assimilation, but without observational constraints (Fig. 1C, D): at each step, a forecasting model produces a deterministic prediction, which is perturbed with calibrated noise and projected back onto the learned climatological manifold through the reverse diffusion process (see "Methods"). This introduces structured stochasticity reflecting the atmosphere's intrinsic variability, while correcting systematic drift that accumulates over extended integration[51]. We evaluate GAP's weather forecasting capability in case studies of high-impact weather events across increasing lead times, and systematically compare its skills against leading forecasting systems. Throughout this section, all GAP forecasts are initialized from GAP's own assimilated analyses in the GAP-OSSE setting, rather than from ERA5 or operational analyses.

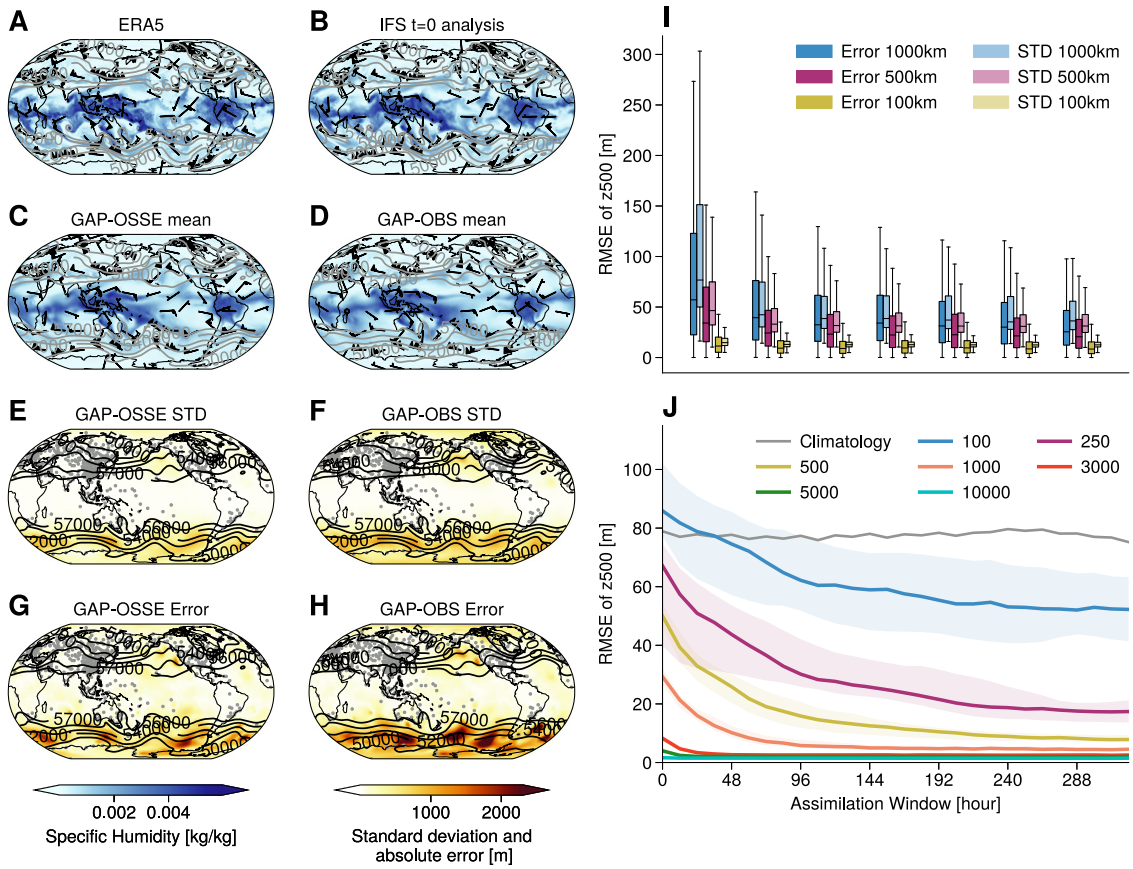


Fig. 3 | Evaluation of GAP's data assimilation capabilities. A–D Spatial patterns of 500 hPa specific humidity and geopotential height at 00 UTC on 3 March 2023 from ERA5 (ground truth), IFS $t = 0$ analysis, GAP-OSSE (perfect observations), and GAP-OBS (real observations). **E–H** Ensemble standard deviation and error patterns at the same time from GAP-OSSE and GAP-OBS, showing strong correspondence between predicted uncertainty (spread) and reference error, indicating reliable uncertainty quantification. **I** RMSE statistics as a function of distance from observation points through the assimilation window, illustrating how error (deep colors) and ensemble variance (shallow colors) decrease near observations and increase with distance. **J** RMSE skill (global mean, withheld from observations) at different observation density levels through the assimilation window for GAP ensemble (colored line with shading), showing progressive acceleration in convergence as observation density increases from climatological levels to 10^4 observational points.

We first examine GAP's ensemble forecasts, initialized from its own assimilated analyses, for four high-impact events spanning lead times from 3 to 42 days: Typhoon In-Fa (3-day lead; Fig. 4A), an atmospheric river making landfall in California (10-day lead; Fig. 4B), a European heatwave (21-day lead; Fig. 4C), and the Madden–Julian Oscillation (MJO; 42-day lead; Fig. 4D). For each case, we present the ERA5 reference, the GAP ensemble mean, ensemble standard deviation, ensemble mean error, and individual ensemble members (additional members in Supplementary Figs. 18–21). For Typhoon In-Fa (Fig. 4A), GAP captures the cyclone's structural evolution in the 500 hPa wind and geopotential height fields at a 3-day lead time. The ensemble mean is naturally smoothed, but individual members exhibit sharp, physically realistic cyclone structures closely aligned with reanalysis — illustrating that GAP generates diverse yet plausible atmospheric realizations rather than blurred averages[51]. The atmospheric river forecast (Fig. 4B) extends to 10 days, where GAP maintains physical consistency in both 850 hPa geopotential height patterns and specific humidity distributions. Atmospheric rivers involve organized, filamentary moisture transport that is sensitive to synoptic-scale flow; the ability to maintain coherent moisture structures at this lead time suggests that the diffusion-based correction helps preserve the organized spatial relationships that purely autoregressive models tend to lose through successive blurring. The European heatwave prediction (Fig. 4C) at 21 days captures both the temperature anomaly pattern and the associated large-scale circulation, indicating that GAP can identify and preserve slow-evolving atmospheric structures that extend predictability beyond conventional deterministic limits. The 42-day MJO forecast (Fig. 4D) shows that GAP represents the characteristic eastward propagation and amplitude modulation of this organized tropical phenomenon over subseasonal timescales — a result that is consistent with the diffusion model's role in maintaining the phase-space structure of large-scale organized modes that would otherwise decay under error accumulation. Across all four cases, the ensemble spread patterns track the spatial structure of the ensemble mean error, extending the spread-skill consistency observed in the assimilation experiments to the prediction setting.

We next conduct a systematic forecast verification over a full calendar year of 2020 (Fig. 5A–C), comparing GAP against forecasting systems representing four distinct modeling strategies: the physics-based operational ensemble (IFS-ENS[36]), deterministic ML models including Pangu[1] and GraphCast[2], physics-ML hybrid NeuralGCM[3], and conditional generative model GenCast[4]. We focus on 500 hPa geopotential height (z500), a standard benchmark for mid-tropospheric weather prediction. A key difference in experimental design is that GAP's forecasts are initialized from its own assimilated analyzes in the GAP-OSSE setting — which uses approximately 5% of grid points as observation locations — whereas all comparison systems are initialized from ERA5 reanalysis or operational analyzes that incorporate far more complete observational coverage. Measured by the Anomaly Correlation Coefficient (ACC, see "Methods", Fig. 5A), GAP performs favorably relative to the deterministic ML models throughout the forecasting period, with skill improvements of approximately 3–14%. This advantage widens beyond day 7, where deterministic autoregressive models decline more rapidly — a pattern consistent with the progressive blurring of spatial structures under repeated deterministic rollout, which GAP's

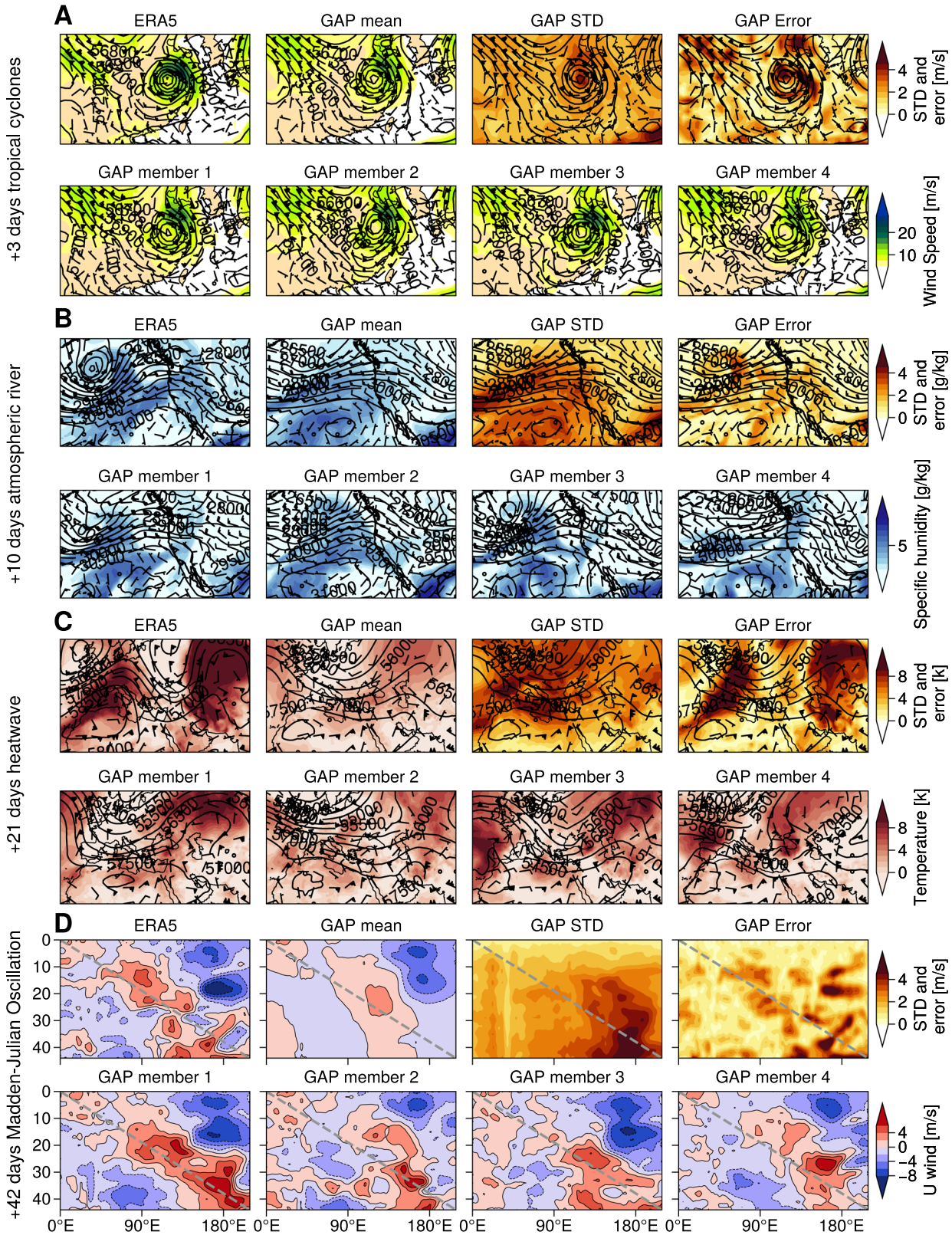


Fig. 4 | GAP produces skillful ensemble forecasts of high-impact weather events across lead times. A 3-day forecast of typhoon In-Fa (July 28, 2021), showing wind speed (color shading) and geopotential height (contours) at 500 hPa from ERA5 reference, GAP ensemble mean, standard deviation, ensemble mean prediction error, and individual ensemble members. **B** 10-day forecast of an atmospheric river hitting California (April 12, 2023), displaying 850 hPa geopotential height (contours) and specific humidity (color shading). **C** 21-day forecast of a European heatwave (June 23, 2023), showing 2-m temperature anomalies (color shading) and 500 hPa geopotential height (contours). **D** 42-day forecast of the Madden–Julian Oscillation (January 20 to March 1, 2023), presenting Hovmöller diagram of zonal wind anomalies at 850 hPa (color shading) across latitudes (0–45°S) and longitudes (0–180°E).

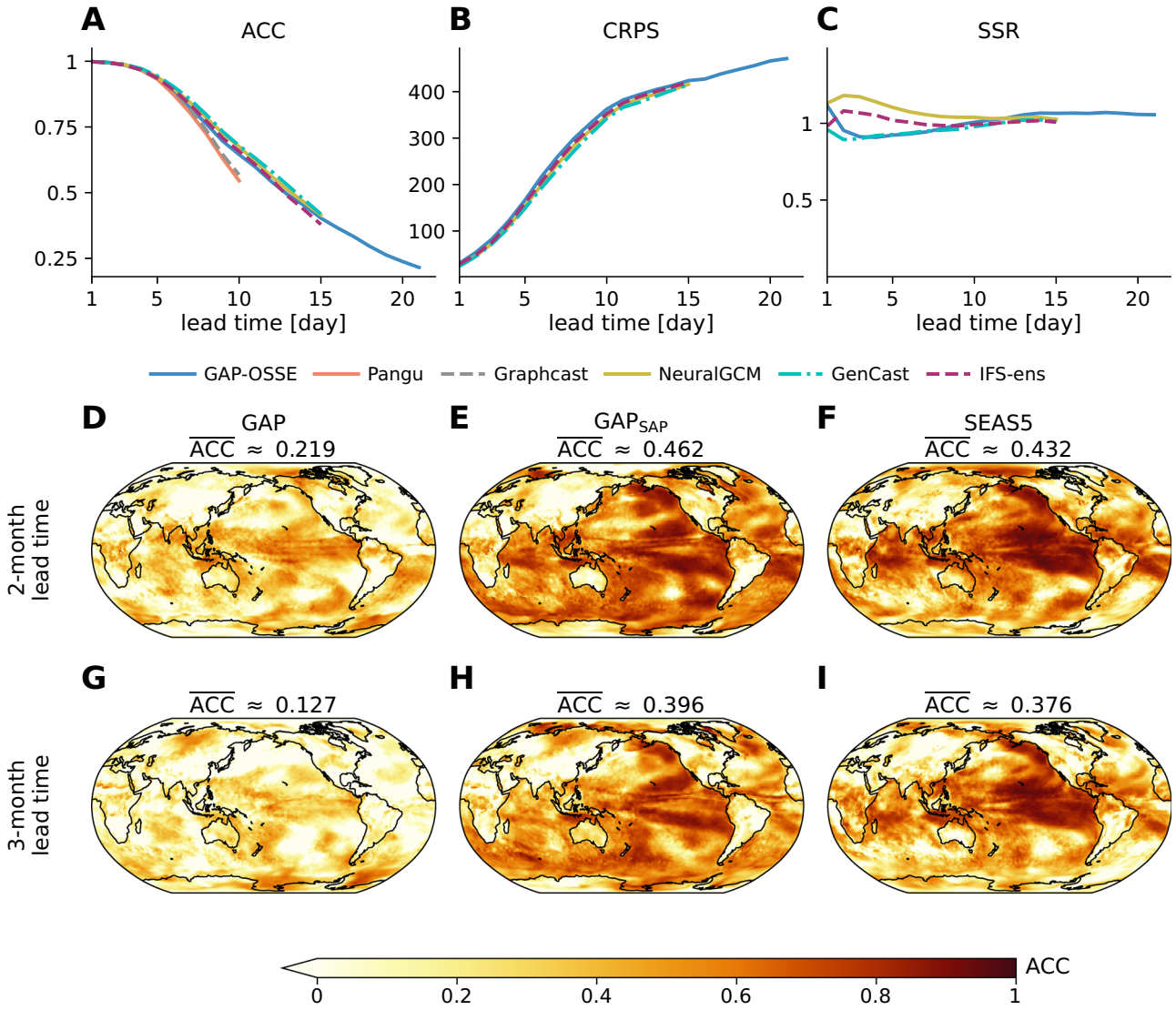


Fig. 5 | Forecast skill of GAP at medium-range and seasonal timescales. **A**–**C** Medium-range forecast skill for 500 hPa geopotential height (z500), comparing GAP-OSSE (initialized from ~5% of the "perfect" synthetic observations) against IFS-ENS, GenCast, NeuralGCM-ENS, Pangu, and GraphCast (all initialized from full analysis fields). Metrics include: **A** anomaly correlation coefficient (deterministic skill); **B** continuous ranked probability score (probabilistic skill); and **C** spread-skill ratio (uncertainty calibration). Seasonal prediction skill for 2-m temperature at 2-month (**D**–**F**) and 3-month (**G**–**I**) lead times. Panels show GAP (**D**, **G**), GAP$_{SAP}$ with simple SST anomaly persistence forcing (**E**, **H**), and ECMWF SEAS5 (**F**, **I**). Latitude-weighted mean ACC values are indicated above each panel.

diffusion-based correction is designed to mitigate. For probabilistic skill, the Continuous Ranked Probability Score (CRPS, see "Methods", Fig. 5B) shows that GAP attains skill comparable to IFS-ENS, GenCast, and NeuralGCM, particularly in the 5–15 day range, where the quality of uncertainty estimates becomes as important as the accuracy of the central forecast. The Spread-Skill Ratio (SSR, see "Methods", Fig. 5C) provides further insight into ensemble calibration. IFS-ENS maintains a ratio close to the ideal value of 1, reflecting decades of careful operational ensemble design. GAP shows a similar tendency through its generative formulation: although its ensemble spread is somewhat tight during the first few forecast days, it gradually converges toward reasonably calibrated uncertainty at longer lead times, consistent with a realistic growth of forecast error. A complementary assessment using central interval coverage and rank histograms (Supplementary Figs. 16 and 17) confirms this trend: the initial mild underdispersion is progressively alleviated at longer leads, with coverage curves moving toward the diagonal and rank histograms approaching uniformity.

These results suggest that GAP's generative assimilation framework can extract sufficient information from limited observational coverage to produce forecasts that are competitive in both deterministic and probabilistic skill with systems that benefit from substantially more complete initialization data. The consistency of performance across deterministic, probabilistic, and calibration metrics — rather than strength in one at the expense of others — points to the value of coupling assimilation and prediction within a shared probabilistic framework.

Potential for seasonal prediction

Atmospheric predictability from initial conditions is inevitably lost within weeks, as chaos amplifies small uncertainties. Beyond this horizon, predictive skill must arise from slowly evolving components of the climate system, primarily the ocean, whose persistent thermal memory constrains atmospheric evolution over months. A central challenge is whether this signal can be properly translated into physically coherent atmospheric responses. GAP's generative framework provides a natural mechanism for

this translation. At each forecast step, the generative model conditions its sampling jointly on the preliminary atmospheric forecast and on imposed oceanic boundary forcing, producing posterior states that are consistent with the predictability source, coherent with the forecasted atmospheric trajectory, and faithful to the multi-scale variability structure learned from reanalysis. This joint conditioning can potentially allow slowly evolving boundary signals to inform the forecast without explicit coupling parameterization, instead relying on co-variability patterns encoded in the learned climatological distribution.

To test this potential for seasonal prediction, we introduce GAP_{SAP} (GAP constrained by Sea Surface Temperature Anomaly Persistence), which incorporates SST information through a deliberately simple procedure (Supplementary Section B.4): initial SST anomalies are persisted throughout the forecasting period, then translated into 2-m air temperature constraints imposed on the generative process. This is not intended as a complete seasonal prediction system — it ignores dynamic ocean evolution, and the SST-to-air-temperature inference introduces additional uncertainty. Rather, it provides a controlled setting for assessing whether the same probabilistic machinery developed for assimilation and weather prediction can incorporate slowly evolving boundary information to extend predictive skill, without burdensome additional training or architectural modification.

To ensure a fair comparison for seasonal prediction, we retrain all models exclusively on pre-2015 data, with the 2015–2024 period held out entirely for evaluation (see "Methods" for the detailed protocol). We evaluate three configurations in a 10-year reforecast experiment: GAP without SST forcing, GAP_{SAP}, and ECMWF's operational seasonal forecasting system SEAS5[56]. The comparison between GAP and GAP_{SAP} (Fig. 5D, E, G, H) isolates the effect of incorporating SST information. GAP_{SAP} yields substantially higher ACC values at both 2-month and 3-month lead times (ACC ≈ 0.462 vs. 0.219 at 2 months; 0.396 vs. 0.127 at 3 months; Supplementary Figs. 25 and 26 for comprehensive comparison). The spatial pattern of improvement is informative: skill gains are most pronounced in regions with strong ocean-atmosphere coupling, suggesting that the diffusion model translates the imposed surface constraints into appropriate regional atmospheric responses rather than simply improving the global mean.

To gauge what this simple configuration achieves relative to an established operational system, we compare GAP_{SAP} with ECMWF's SEAS5 (Fig. 5F, I). At a global level, GAP_{SAP} achieves higher latitude-weighted ACC (0.462 vs. 0.432 at 2 months; 0.396 vs. 0.376 at 3 months), but with distinct regional variation (Supplementary Fig. 26). Over ocean-dominated regions such as the North Pacific, GAP_{SAP} frequently exceeds SEAS5, consistent with effective propagation of the prescribed oceanic constraint into the overlying atmosphere. Over continental interiors such as Africa, skill is weaker — likely reflecting compound uncertainties from SST persistence errors, the SST-to-air-temperature inference, and the difficulty of propagating oceanic information over land through the inpainting-based mechanism.

These results provide initial evidence that GAP's conditioning mechanism generalizes from assimilation to seasonal timescales. The simplicity of the current SST approach also points to clear avenues for improvement — including dynamical ocean prediction and more sophisticated coupling strategies — which we expect would strengthen skill, particularly over land regions.

Climate simulation

The seasonal prediction experiments demonstrate that GAP can incorporate boundary forcing to extend skill beyond the weather timescale. A natural question is whether this capability extends further — to climate timescales where initial atmospheric state information has been entirely dissipated, and the system's behavior is governed by internal variability and external forcing alone. This regime poses a distinct challenge for autoregressive ML models, which typically accumulate errors over extended integration, leading to drift or numerical instability[8,9,25,26]. For GAP, the generative posterior sampling approach that projects each forecast step back onto the learned climatological manifold should, in principle, prevent such drift. We test this through three experiments: a free-running simulation without any external constraint, to assess whether GAP maintains numerical stability and realistic variability over extended periods; a historical simulation with prescribed sea-air temperatures, to evaluate whether the framework produces appropriate atmospheric responses to observed oceanic forcing; and a set of boundary-forced experiments using sea-air temperatures from CESM2 climate projections under different Shared Socioeconomic Pathways, to examine whether GAP's atmospheric model produces physically reasonable responses when driven by boundary conditions outside its training distribution.

Our millennium-scale free-running simulation provides the most basic test: without any external constraint, GAP maintains physically realistic temperature variability around a stable mean for over 1000 years (Fig. 6A), avoiding the drift that typically limits ML-based forecasting models. This stability extends to the seasonal cycle, which GAP reproduces with correct amplitude and phase even after a millennium of integration (Fig. 6B). Beyond stability, the free-running model reproduces several key modes of climate variability: realistic tropical cyclone spatial distributions and intensity characteristics (Fig. 6C), mid-latitude blocking patterns (Fig. 6D), and the Arctic Oscillation (Fig. 6E). These emerge without being explicitly trained for, suggesting that the learned climatological distribution encodes sufficient dynamical structure to sustain these phenomena through internal variability alone. The free-running model does, however, show limitations for phenomena strongly tied to ocean forcing — notably the Southern Oscillation Index (Fig. 6F), which requires ocean-atmosphere interaction that a free-running atmospheric model cannot represent.

When provided with prescribed historical sea-air temperatures as boundary forcing (GAP_{PSAT}), the model successfully reproduces these ocean-driven patterns (Fig. 6F, bottom row), confirming that the limitation lies in the absence of oceanic information rather than in the atmospheric framework itself. Further temporal consistency analyzes such as autocorrelation, temporal spectra, persistence of circulation regimes and blocking duration, are shown in Supplementary Fig. 29. To examine whether GAP produces physically reasonable atmospheric responses under boundary conditions not encountered during training, we drive it with sea-air temperatures from CESM2 simulations under three Shared Socioeconomic Pathways (SSP1-2.6, SSP3-7.0, SSP5-8.5). Under these externally prescribed conditions, GAP maintains stable atmospheric structure and circulation patterns (Supplementary Figs. 32–35), reproduces key modes of variability consistent with the imposed forcing (Supplementary Figs. 37–40), and captures the expected trends in temperature profiles, though it exhibits distributional shift problems common to other ML models (Supplementary Fig. 41). It is worth noticing that these experiments should not be interpreted as independent climate projections — the atmospheric response is entirely conditioned on boundary information produced by an external coupled model — but they suggest that GAP's atmospheric framework can respond appropriately to forcing conditions.

We also compare the climatological bias of GAP's simulations over the period 1959-1978 against CMIP6 models[13] and ACE2-ERA5[25], an ML atmospheric emulator trained on ERA5 (Fig. 6G). We examine three GAP configurations with increasing levels of constraint: unconditional generation (GAP_{clim}), a free-running simulation initialized in 1959 (GAP), and the boundary-forced simulation (GAP_{PSAT}). GAP reduces climatological biases relative to most CMIP6 models, and GAP_{PSAT} achieves further improvement, approaching the performance of ACE2-ERA5 for most variables while still falling short for some, such as z500.

These results suggest the GAP framework provides the stability needed for extended atmospheric integrations. The conditioning mechanism demonstrated for seasonal prediction extends naturally to climate-scale boundary forcing, enabling GAP to serve as an efficient atmospheric model for exploring atmospheric responses to prescribed oceanic and radiative conditions. We note, however, that the future-scenario experiments rely on boundary conditions derived from CESM2 and should therefore be

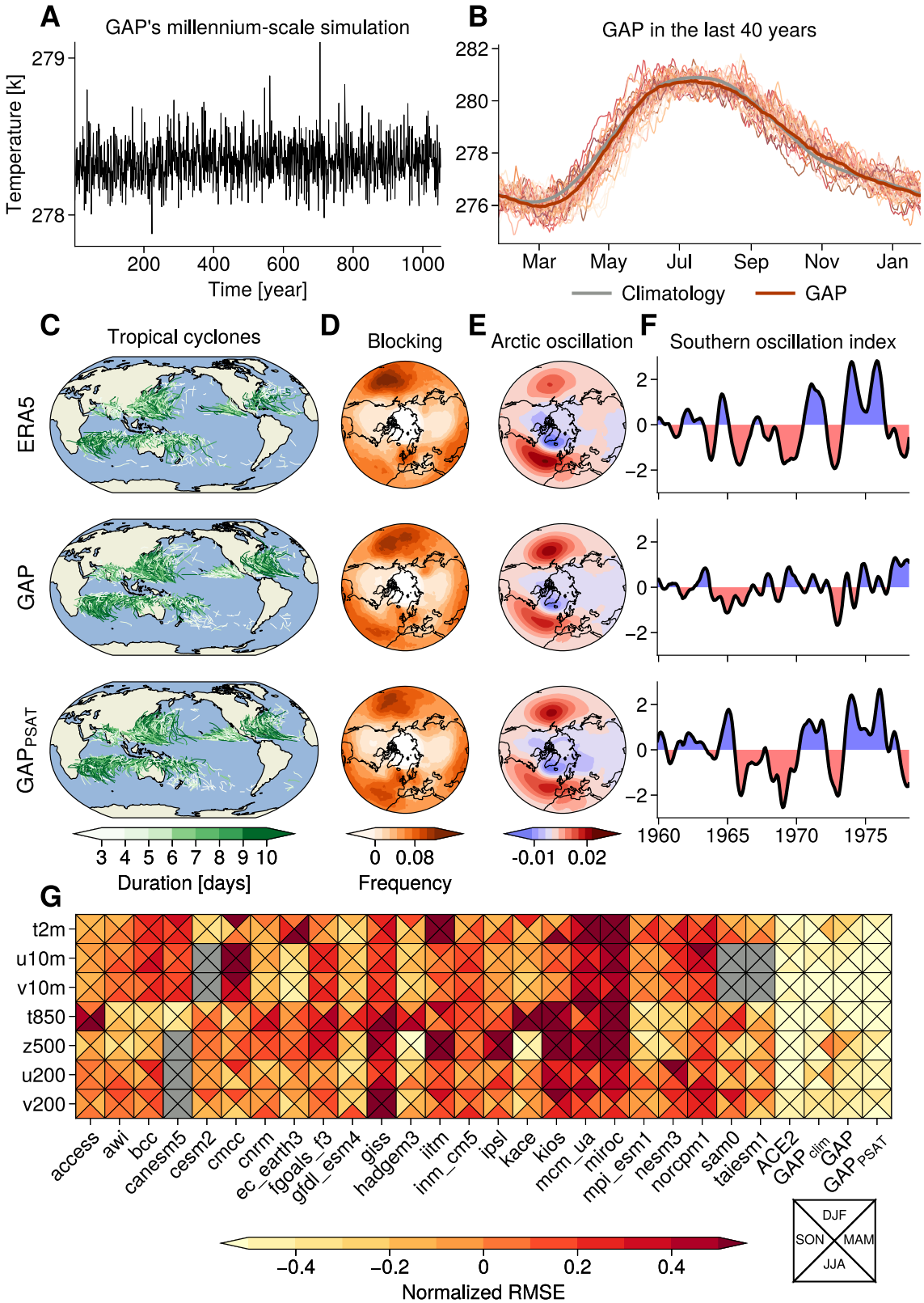


Fig. 6 | GAP sustains stable long-term climate simulation and reproduces key modes of variability. A Millennial-scale free simulation: global mean surface air temperature from a run of more than 1,000 years (initialized in 1980 and integrated to year 3030). **B** Reproduction of the seasonal cycle in the final 40 years (2990–3030), with individual years (thin red lines) and climatological mean (thick red line) compared against ERA5 climatology (blue line). **C**-**F** Representation of key modes of climate variability during the test period 1959-1978 in ERA5 (top row), GAP free run (middle row), and GAP with prescribed sea air temperatures (GAP_{PSAT}, bottom row): **C** tropical cyclone track density with duration indicated by color; **D** mid-latitude blocking frequency; **E** Arctic Oscillation pattern represented by sea-level pressure regression; **F** Southern Oscillation Index time series illustrating interannual variability. **G** Spatial RMSE of global seasonal climatologies across selected CMIP6 models, ACE2, and GAP (GAP_{clim}, unconditional GAP generation) over the same period (1959–1978), normalized by the median RMSE (lower values indicate better performance).

interpreted as conditional atmospheric responses to externally prescribed forcing rather than as independent climate projections.

Connection to existing approaches

The application of ML to weather and climate science is advancing rapidly, with new methods appearing across data assimilation[23,40,41], weather forecasting[1–4], subseasonal-to-seasonal prediction[8,57], and climate simulation[25]. Several recent systems already couple learned forecasting models with data assimilation within a single framework[7,23,58]. To clarify GAP's contribution, we discuss how it relates to existing approaches along the assimilation and prediction axes.

ML-based data assimilation methods can be organized along two classical paradigms — variational and sequential — distinguished by whether they seek a single optimal state estimate or propagate a state distribution forward in time.

Variational data assimilation finds the atmospheric state that best reconciles background information with observations, typically by minimizing a cost function. ML has entered this paradigm in two main ways. *Auto-differentiation–based methods* replace the adjoint model in 3D/4D-Var with automatic differentiation through data-driven forecasting models, greatly accelerating gradient computation[23,59]. However, because the underlying forecast models operate at fixed time steps, these methods cannot assimilate observations arriving at higher temporal frequency. *End-to-end supervised methods* frame assimilation as a regression from background states and observations to a reference analysis, using architectures such as masked transformers[60] and neural processes[7]. While some of these methods do provide uncertainty estimates, the quantification is inherently tied to the fixed observation–reanalysis pairing seen during training: the learned uncertainty reflects the residual statistics of this regression rather than the actual uncertainty arising from the background quality, the observation network configuration, or the flow regime. When observation coverage or error characteristics change at inference time as they inevitably do in operational settings, these methods offer no principled mechanism to adapt their uncertainty estimates accordingly.

Sequential assimilation propagates a state distribution forward in time and updates it as new observations arrive, separating a *prediction step* from an *analysis step*. Recent work uses deep generative models — particularly diffusion models — to represent and sample from the complex, high-dimensional distributions involved. Existing methods differ in what distribution they learn and how they adapt it to observations.

One line of work learns the state transition distribution $p(\mathbf{X}_{t+\Delta t}|\mathbf{X}_t)$ from reanalysis data, then incorporates observations during sampling to approximate the posterior $p(\mathbf{X}_{t+\Delta t}|\mathbf{X}_t, \mathbf{O}_{t+\Delta t})$ — for example via diffusion posterior sampling[40] or conditional score modeling[41]. A key limitation is that the transition distribution is shaped by reanalysis backgrounds, which are high-quality, dynamically balanced states produced by mature operational systems. In practice, however, the background $\mathbf{X}_t$ is a forecast that may carry substantial, flow-dependent biases — a setting that diverges from the stable, information-rich conditions under which the distribution was trained. This mismatch can degrade both the state estimates and the associated uncertainty quantification.

An alternative line of work addresses this gap by using sequential Monte Carlo frameworks[22] or fine-tuning[24] the generative model during the assimilation cycle, allowing the distribution to adapt to the evolving background. While conceptually appealing, continuously updating a deep generative model in real time introduces significant computational overhead and risks distributional instability — both of which pose challenges for operational deployment where reliability and speed are essential.

GAP takes a distinct path from both paradigms. Rather than learning a state transition distribution $p(\mathbf{X}_{t+\Delta t}|\mathbf{X}_t)$ shaped by reanalysis backgrounds, or continuously adapting a generative model during operation, GAP pre-learns the *climatological distribution* of atmospheric states — a general, task-independent prior that reflects the climate system's intrinsic variability rather than the systematic characteristics of any particular forecast model or analysis system. At inference time, this fixed prior is conditioned on observational and predictive constraints through sequential Bayesian updates.

This design offers three advantages. First, because the prior captures the climate system's own variability rather than the residual error of a specific model, it provides uncertainty estimates that remain physically grounded as observation networks and forecast models change. Second, because the distribution is learned once and held fixed during operation, GAP avoids the computational cost and potential instabilities of online distributional adaptation. Third, and most fundamentally, the same learned climatological distribution serves as the foundation for prediction and climate simulation — establishing a bidirectional link between assimilation and forecasting that is unavailable to methods designed for a single task. We discuss this connection next.

ML weather models now match or exceed numerical weather prediction accuracy at greatly reduced cost[1,2], but extending these capabilities requires solving two additional problems: uncertainty quantification and long-term stability.

For uncertainty, methods have progressed from ad hoc perturbations[1] through distribution-aware training losses such as CRPS[57,61] to conditional generative models that capture structured forecast-error distributions[4]. For stability, strategies include physics-constrained architectures[3], multi-step autoregressive training[8], resolution reduction[9], and prescribed boundary conditions[25]. Each entails trade-offs. Physics-constrained designs add complexity; spatially averaged probabilistic losses improve calibration but do not enforce spectral consistency of the predicted fields; conditional generative models like GenCast[4] capture structural uncertainty but have not demonstrated stable long-term integration, likely because they model forecast-error distributions rather than the climatological distribution; and resolution reduction or boundary conditioning extends forecast horizons at the cost of spatial detail or coupled dynamics. In practice, most methods are optimized for specific timescales, making seamless prediction across weather and climate regimes difficult within a single framework.

GAP addresses both problems through a single mechanism. At every forecast step, the learned climatological distribution generates calibrated ensemble perturbations reflecting how uncertainties evolve, while simultaneously preventing the predicted state from drifting outside the physically plausible region of phase space. Because this mechanism is timescale-agnostic — operating identically whether the lead time is hours or centuries — it connects ensemble weather forecasting with long-term climate stability without separate training or architectural modification.

Discussion

Machine learning is rapidly reshaping how we simulate and predict the climate system, delivering forecast skill that rivals or exceeds decades of numerical model development[1–4]. Yet the advances have arrived piecemeal: data assimilation, weather forecasting, seasonal prediction, and climate simulation are often each addressed by dedicated architectures, training objectives, and operational pipelines[45,62], with limited awareness that they are manifestations of a single underlying challenge. Stepping back, a common thread runs through all of them: each seeks to estimate the distribution of atmospheric states under a particular combination of observational, predictive, and forcing constraints. Assimilation narrows the distribution using observations; forecasting evolves it forward under chaotic dynamics; seasonal prediction sustains it through slowly varying boundary forcing; and climate simulation tracks how external forcings reshape it over decades and centuries. The differences among these tasks lie not in the object being estimated, but in which constraints are active and over what timescales.

This recognition motivates a simple but consequential design choice: learn the climatological distribution once, and condition it flexibly for each task. The result is a bidirectional link between assimilation and prediction. In one direction, the climatological prior constrains analyzes to physically plausible, uncertainty-aware initial states. In the other direction, the same prior anchors long-range forecasts against drift, projecting each prediction step back onto the climatological distribution rather than allowing autoregressive errors to compound unchecked. This creates a self-reinforcing

cycle: a better-learned climatology yields sharper analyzes, which in turn produce better-initialized forecasts, whose statistics further validate and refine the climatological representation.

The strength of this unified framework becomes evident not at any single task, but across the full assimilation–prediction continuum. When conditioned on sparse observations, this learned distribution produces assimilated states with reasonably calibrated uncertainty estimates, where ensemble spread tracks the spatial structure of analysis error. These analyzes, derived from approximately 5% of grid-point coverage, support ensemble weather forecasts competitive in probabilistic skill with leading systems initialized from far more complete observational data. At seasonal timescales, the same conditioning mechanism incorporates slowly evolving oceanic boundary information to extend predictive skill. And over millennial-scale free-running integrations, the diffusion-based correction maintains numerical stability and realistic atmospheric variability — including tropical cyclones, blocking events, and the Arctic Oscillation. We note that the future-scenario experiments rely on boundary conditions derived from CESM2 and should be interpreted as conditional atmospheric responses to prescribed forcing rather than as independent climate projections.

Several directions could substantially strengthen this framework. GAP is an approximate Bayesian realization of PAP: observational conditioning is implemented through inpainting and diffusion posterior sampling, predictive conditioning through SDEdit[49], and forcing information enters only indirectly through prescribed surface air temperature. More advanced posterior-sampling schemes[47,63] would tighten the gap between generated samples and the target conditional distributions, improving ensemble calibration, particularly at short lead times where GAP currently exhibits mild underdispersion. Incorporating greenhouse gas concentrations and aerosol loadings as explicit forcing constraints would free GAP from its current reliance on externally prescribed sea–air temperatures, enabling it to represent radiative forcing pathways directly. Extending the observation operator beyond grid-point measurements to accommodate satellite radiances and other remote-sensing data would bring the assimilation substantially closer to operational requirements. Finally, the modular architecture of GAP, where the generative model and the forecasting model are independent, interchangeable components, means that advances in either can be incorporated without retraining the other. Replacing the current data-driven forecasting model with a physics-based dynamical core, for instance, would combine GAP's probabilistic framework with explicit conservation laws, potentially offering the physical guarantees of numerical models together with the distributional flexibility of generative modeling.

More broadly, GAP reveals a connection between generative modeling and the dynamics of chaotic systems. Both transform unstructured inputs into structured patterns: a diffusion model maps Gaussian noise to atmospheric states through a learned score function, while the climate system maps arbitrary perturbations to its characteristic variability through nonlinear dynamics. The difference is that generative models learn this transformation from data, while chaotic systems realize it through physical dynamics. The climate system, with its well-characterized chaotic behavior, rigorous prediction metrics, and century-long observational record, provides a more demanding test for generative models than conventional benchmarks[64] — one where physical consistency, spectral fidelity, and calibrated uncertainty matter far more than perceptual quality. That a generative model can approximate the atmospheric distribution well enough to sustain realistic variability over a millennium suggests possibilities for productive exchange among chaos theory, machine learning, and climate science.

Methods

Probabilistic assimilation and prediction

We tackle the task of assimilation and prediction in a unified formulation, i.e., estimating the probabilistic distribution of atmospheric states given observational, predictive, and external forcing constraints.

Let $\mathbf{X} \in \mathbb{R}^d$ represent the atmospheric state vector, which comprises five atmospheric variables (zonal and meridional wind velocity, temperature, specific humidity, and geopotential height) defined on a three-dimensional grid (latitude × longitude × pressure levels), along with four surface variables (10-m winds, 2-m temperature, and sea-level pressure) defined over latitude and longitude spans (Supplementary Table 1). At our current resolution of $1^\circ \times 1^\circ$ with 13 pressure levels, the state space dimension $d \approx 4.5$ million.

We now define three key timestamps that delineate different phases in an assimilation and prediction cycle:

- t_a: The start of the assimilation window, when observations begin to constrain state uncertainty.
- t_0: The transition point between assimilation and prediction, after which no new observations are available.
- t_p: The end of the prediction window, which may extend to climate timescales.

Given this setup, we address the following three interconnected estimation tasks:

1. *Climatological distribution estimation*: For $t < t_a$, we estimate the baseline climatological distribution by maximizing

$$p_{\theta^*} = \arg\max_{p_\theta} \prod_{i=1}^{N} p_\theta(\mathbf{X}_i|\Theta), \tag{1}$$

where p_θ is a parameterized probability distribution with parameters θ, $\mathbf{X}_{i\in[1,\ N]}$ are atmospheric state samples from climate reanalysis, and Θ represents external forcing conditions.

2. *Assimilation*: For $t = t_a$, given observations $\mathbf{O}_{t_a}$ and an observation operator defining $p(\mathbf{O}|\mathbf{X})$, we estimate the conditional distribution

$$p(\mathbf{X}_{t_a}|\mathbf{O}_{t_a}) \propto p_{\theta^*}(\mathbf{X}_{t_a}|\Theta)p(\mathbf{O}_{t_a}|\mathbf{X}_{t_a}) \tag{2}$$

For $t \in [t_a + \Delta t, t_0]$, where Δt denotes the computation time step, given a sequence of observations $\mathbf{O}_{t_a:t}$, and a forecasting model $\mathbf{M}$ that induces transition $p(\mathbf{X}_{t+\Delta t}|\mathbf{X}_t)$, we then estimate the evolving state distribution through approximate recursive Bayesian updates:

$$p(\mathbf{X}_t|\mathbf{O}_{t_a:t}) \approx \int p(\mathbf{X}_t|\mathbf{X}_{t-\Delta t}, \mathbf{O}_t)p(\mathbf{X}_{t-\Delta t}|\mathbf{O}_{t_a:t-\Delta t})d\mathbf{X}_{t-\Delta t} \tag{3}$$

3. *Prediction:* For $t > t_0$, starting from the assimilated distribution $p(\mathbf{X}_{t_0}|\mathbf{O}_{t_a:t_0})$, project forward to estimate:

$$p(\mathbf{X}_{t_0:t_p}|\mathbf{X}_{t_0}, \Theta) = \prod_{t=t_0}^{t_p-\Delta t} p(\mathbf{X}_{t+\Delta t}|\mathbf{X}_t, \Theta) \tag{4}$$

This prediction spans multiple characteristic ranges from weather to climate, with dominant source of predictability shifting from the initial state $\mathbf{X}_{t_0}$ to the external forcing Θ.

These three tasks are linked through their dependence on the underlying climatological distribution, with each task addressing distinct sources of uncertainty — from observational uncertainties in data assimilation, to chaos-induced growth of forecast uncertainty in prediction, to forcing-dependent variability in long-term climate projections. Together, they constitute the PAP framework.

Generative assimilation and prediction

GAP offers a practical solution to the PAP formulation regarding the computation of the distributions $p_{\theta^*}(\mathbf{X}|\Theta)$, $p(\mathbf{X}_{t_a}|\mathbf{O}_{t_a})$, $p(\mathbf{X}_t|\mathbf{X}_{t-\Delta t}, \mathbf{O}_t)$, $p(\mathbf{X}_t|\mathbf{X}_{t-\Delta t})$, and $p(\mathbf{X}_t|\mathbf{X}_{t-\Delta t}, \Theta)$, as explained in detail below. An ablation study of GAP can be found in Supplementary Section D.7.

$p_{\theta^*}(\mathbf{X}|\Theta)$: climatological distribution estimation

To approximate the climatological distribution of atmospheric states, we parameterize $p_{\theta^*}(\mathbf{X}|\Theta)$ using a probabilistic diffusion model trained on atmospheric reanalysis data. Specifically, the diffusion framework applies a forward process defined by the following stochastic differential equation (SDE):

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, \tau)d\tau + g(\tau)d\mathbf{w}, \quad \tau \in [0, T] \tag{5}$$

where $\mathbf{w}$ is a standard Wiener process,$\mathbf{f}$ is the drift term, and g is the diffusion term. $\mathbf{f}$ and g are predetermined so that the forward SDE transforms the target climatological distribution to a standard Gaussian distribution: $p(\mathbf{x}_{\tau=0}) = p_{\theta^*}(\mathbf{X}|\Theta)$, $p(\mathbf{x}_{\tau=T}) = \mathcal{N}(\mathbf{0}, \mathrm{I})$.

The corresponding reverse SDE for generation is given by

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, \tau) - g^2(\tau)\nabla_{\mathbf{x}} \log p(\mathbf{x})]d\tau + g(\tau)d\bar{\mathbf{w}} \tag{6}$$

where $\bar{\mathbf{w}}$ is a standard Wiener process running backward in time, $\nabla_{\mathbf{x}} \log p(\mathbf{x})$, known as the score function, is approximated using a deep neural network s_θ trained for score matching:

$$\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau) \approx \arg\min_{s_\theta} \mathbb{E}_{\tau,\mathbf{x}_0} \mathbb{E}_{\mathbf{x}_\tau|\mathbf{x}_0} \| s_\theta(\mathbf{x}_\tau, \tau) - \nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau|, \mathbf{x}_0)\|_2^2 \tag{7}$$

We parameterize s_θ with a U-Net backbone[39], extended with residual blocks at intermediate scales, and time-step embeddings to effectively capture both spatial hierarchies and temporal noise levels. For training, we employ the Adam optimizer with learning rate 1×10^{-4}, batch size of 4, and train for over 1000 epochs until convergence. A cosine noise schedule is used to ensure smooth transitions across noise levels, and gradient clipping is applied to stabilize optimization. Training is performed on NVIDIA A100 GPUs, with mixed precision enabled for efficiency. For generation, we adopt denoising diffusion implicit model sampling, which significantly reduces the number of function evaluations required for high-fidelity output[65]. Additional details on data preprocessing, model hyperparameters, noise scheduling, and training implementation are provided in Supplementary Section B.1.

$p(\mathbf{X}_t|\mathbf{O}_t)$: enforcing observational constraint

At the start of assimilation, we combine the learned climatological prior with observational evidence to obtain posterior state estimates. For notational clarity, we drop the subscript t_a from $\mathbf{X}_{t_a}$ and $\mathbf{O}_{t_a}$ when discussing the diffusion process, and use subscripts to denote diffusion time instead.

The general measurement process can be formulated as

$$\mathbf{O} = \mathcal{A}(\mathbf{x}_0) + \mathbf{n}, \qquad \mathbf{O}, \mathbf{n} \in \mathbb{R}^k, \ \mathbf{x}_0 \in \mathbb{R}^d \tag{8}$$

where $\mathcal{A}(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^k$ is the forward measurement operator and $\mathbf{n}$ denotes measurement noise. To incorporate observational constraints into the generative process, we modify the learned score function to include observational information:

$$\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau|, \mathbf{O}) = \nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau) + \nabla_{\mathbf{x}_\tau} \log p(\mathbf{O}|, \mathbf{x}_\tau) \tag{9}$$

The first term is provided by the score function of the pretrained diffusion model. The second term represents the likelihood gradient, which is not tractable in closed form because the observations $\mathbf{O}$ depend only on the clean state x_0 at diffusion time $\tau = 0$, but not directly on the noisy intermediate state $\mathbf{x}_\tau$.

Following[47,48], we assume that the clean state x_0 can be approximated from the intermediate state $\mathbf{x}_\tau$, and replace the intractable likelihood $p(\mathbf{O}|\mathbf{x}_\tau)$ by $p(\mathbf{O}|\hat{\mathbf{x}}_0)$, where $\hat{\mathbf{x}}_0$ denotes the estimated clean state obtained either via Tweedie's formula[47] or a one-step approximation[48]. For instance, under Gaussian measurement noise $\mathbf{n} \sim \mathcal{N}(0, \sigma^2 I)$ with measurement operator $\mathcal{A}(\cdot)$, the likelihood term becomes

$$\nabla_{\mathbf{x}_\tau} \log p(\mathbf{O}|, \mathbf{x}_\tau) \approx -\frac{1}{\sigma^2} \nabla_{\mathbf{x}_\tau} \| \mathbf{O} - \mathcal{A}(\hat{\mathbf{x}}_0)\|_2^2 \tag{10}$$

and the posterior score is then obtained by adding this correction to the diffusion prior score. In the OSSE experiments, synthetic observations are generated by perturbing the ERA5 reference fields at designated observation locations with Gaussian noise $\mathbf{n} \sim \mathcal{N}(0, \sigma^2 I)$, where $\sigma = 0.1$ (in units of the climatological standard deviation of each variable).

For linear $\mathcal{A}(\mathbf{x}) = A\mathbf{x}$, the likelihood gradient can be evaluated directly under the clean-state approximation; for nonlinear but differentiable $\mathcal{A}$, we backpropagate through the Jacobian $J_{\mathcal{A}}(\hat{\mathbf{x}}_0)$; if $\mathcal{A}$ is non-differentiable, we use a smooth or learned differentiable surrogate and the same update applies[47]. These updates remain approximate because the diffusion-time likelihood is evaluated through the estimated clean state.

To further encourage coherence between observed and unobserved regions, we adopt the "time-travel" strategy, in which the sampling trajectory is periodically perturbed by reverting to earlier diffusion steps. This resampling mechanism improves global consistency across the reconstructed fields. The complete observation-error specification and all likelihood and guidance weights are detailed in Supplementary Sections A.3 and B.3, respectively.

$p(\mathbf{X}_t|\mathbf{X}_{t-\Delta t}, \mathbf{O}_t)$ & $p(\mathbf{X}_t|\mathbf{X}_{t-\Delta t})$: enforcing predictive constraints

To incorporate predictive information into the generative process, GAP combines the learned climatological prior with forecasts from a predictive model $\mathbf{M}$ that induces state transition $p(\mathbf{X}_t|\mathbf{X}_{t-\Delta t}, \mathbf{O}_t)$ (during assimilation) and $p(\mathbf{X}_t|\mathbf{X}_{t-\Delta t})$ (during prediction). While various models could serve this purpose, including physics-based approaches, we demonstrate GAP's capabilities using a data-driven forecasting model.

The predictive model $\mathbf{M}$ maps the current atmospheric state to the next state that is 6 h ahead. This model, which we call ConvCast, is trained on 39 years of ERA5 reanalysis data (1979–2017) at 1° spatial resolution, minimizing a weighted mean squared error loss function aligned with GraphCast's weighting scheme[2]. The training process incorporates cosine-annealing learning rate scheduling initialized at 0.001 and utilizes a scheduled drop path with a 0.2 ratio to prevent overfitting. This combination of architectural design and training strategies ensures robust performance in weather prediction tasks while maintaining the computational efficiency needed for operational forecasting applications. Supplementary Section B.2 provides details of this predictive model.

Since $\mathbf{M}$ inevitably contains systematic biases and uncertainties, we adopt Stochastic Differential Equation Editing (SDEdit)[49] to enforce predictive constraints. Rather than generating samples from pure noise, SDEdit initializes the diffusion process with the raw forecast $\mathbf{M}(\mathbf{X}_t)$ perturbed by calibrated noise: the forward SDE is run from $\tau = 0$ to a chosen noise level τ^*, and the reverse SDE is then solved back to $\tau = 0$ to yield the probabilistic estimate $p(\mathbf{M}(\mathbf{X}_t))$.

The parameter τ^* governs the trade-off between relying on raw forecasts and reverting to the climatological prior. In practice, τ^* is selected via spectral diagnostics and probabilistic forecast scores (e.g., CRPS).

For assimilation steps $p(\mathbf{X}_t|\mathbf{X}_{t-\Delta t}, \mathbf{O}_t)$, the same strategy is used while additionally enforcing observation likelihoods during reverse sampling. Further practical guidance for selecting τ^* can be found in Supplementary Section B.6.

$p(\mathbf{X}_t|\mathbf{X}_{t-\Delta t}, \Theta)$: enforcing forcing constraints

To represent how forcings (Θ) influence predictions, we capture their impacts through key state variables, primarily focusing on ocean temperatures. Since oceans absorb over 90% of excess greenhouse gas heat, they contain the most salient climate change signals. Our approach incorporates the ocean temperature constraints into the generative process by adding a modified likelihood term analogous to that used for observational

constraints, effectively translating ocean climate signals into appropriate weather patterns.

For seasonal prediction, we first project future sea surface temperatures using anomaly persistence, then infer corresponding 2-m air temperatures, and finally incorporate these constraints throughout simulations (see Supplementary Section B.4 for implementation details). To ensure a strictly out-of-sample evaluation, all trainable components, including the diffusion-based GAP model, the ConvCast forecasting model, and the SST-to-t2m statistical mapping, are retrained exclusively on ERA5 data from 1979 to 2014. The entire 2015–2024 period is held out. Under this protocol, for each evaluation year within 2015–2024, seasonal hindcasts are initialized in April, May, and June, respectively, and integrated forward to produce seasonal forecasts.

For historical climate simulations, we incorporate sea-air temperatures from ERA5 as boundary forcings. In climate projections, we incorporate sea-air temperatures from CESM2 simulations under different forcing scenarios (SSP1-2.6, SSP3-7.0, SSP5-8.5, see Supplementary Section B.5).

While this approach intentionally ignores dynamic ocean-atmosphere feedbacks, it enables GAP to bridge weather prediction and climate projection while maintaining physical consistency across timescales without additional training. We recognize the limitations of this approach — particularly the lack of dynamic ocean-atmosphere feedback and reliance on external models to represent climate change signals. Future work will aim to directly represent forcing-response relationships in a coupled modeling framework.

Data

We train the generative model and the predictive model using hourly atmospheric reanalysis data from ECMWF's ERA5 reanalysis at 1° spatial resolution (Supplementary Section A.1). The training dataset spans from 1979 to 2017 and includes nine key weather variables across 13 standard pressure levels (50–1000 hPa). For model evaluation across different timescales, we use several reference datasets:

- In data assimilation experiments, we use observational data from the Integrated Global Radiosonde Archive (IGRA) and Global Hourly Integrated Surface Database (ISD).
- For weather forecasting (1–14 days), we benchmark against ECMWF's operational ensemble forecast system (IFS-ENS) as well as recently introduced ML models.
- For seasonal prediction (1–3 months), we compare with ECMWF's fifth-generation seasonal forecasting system (SEAS5).
- For climate simulation and projection experiments, we utilize sea-air temperatures from CESM2 simulations under various shared socioeconomic pathway scenarios (SSP1-2.6, SSP3-7.0, SSP5-8.5).

For operational observations, we employ two primary datasets: the Integrated Global Radiosonde Archive (IGRA) and the Global Hourly Integrated Surface Database (ISD). IGRA provides quality-controlled radiosonde observations from over 2800 globally distributed stations, offering vertical profiles of atmospheric variables since 1905, while ISD integrates hourly surface observations from more than 100 sources, measuring parameters including wind, temperature, pressure, and precipitation (Supplementary Section A.2). Extensive preprocessing was applied to handle missing values and measurement errors, with specific protocols for near-surface wind components, surface air temperature, pressure, and radiosonde data. All grid-based data underwent first-order conservative regridding following established benchmarking protocols (Supplementary Section A.3).

Statistical methods

We implement a comprehensive evaluation framework that spans deterministic, probabilistic, and physical metrics across different spatial and temporal scales. For deterministic assessment (Supplementary Section C.1), we use:

- The Root Mean Square Error (RMSE) to quantify the magnitude of forecast errors.
- The ACC to measure pattern similarity between forecast and observed anomalies.

For probabilistic evaluation of ensemble forecasts (Supplementary Section C.2), we employ:

- The CRPS to assess the overall quality of probabilistic forecasts.
- The SSR to evaluate ensemble reliability by comparing ensemble spread with forecast error.

To assess physical consistency and scale-dependent behavior, we analyze

- spectral characteristics using zonal Fourier analysis.
- statistical distribution matching using the Kolmogorov–Smirnov test at individual grid points (Supplementary Section C.4).
- Multi-variable consistency through derived fields including wind speed, mass streamfunction, geostrophic/ageostrophic wind separation, humidity parameters, and water content (Supplementary Section C.5).

For high-impact weather and climate phenomena, we implement specialized detection algorithms for tropical cyclones, atmospheric blocking and major oscillation indices (Arctic and Southern Oscillation, Supplementary Section C.7).

Data availability

All datasets used in this study are publicly available through their respective repositories. The ERA5 reanalysis data were obtained from the Copernicus Climate Change Service's Climate Data Store (CDS) (https://cds.climate.copernicus.eu). The ECMWF ensemble (ENS) and high-resolution (HRES) forecast datasets were accessed from https://apps.ecmwf.int/datasets/data/tigge/. For the quickest access to these datasets, the WeatherBench2 data archive provides an efficient alternative (https://console.cloud.google.com/storage/browser/weatherbench2). The Integrated Surface Database (ISD) was obtained from the National Centre for Environment Information (NCEI) (https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database), while the Integrated Global Radiosonde Archive (IGRA) sounding data were accessed through the NCEI Weather Balloon Database (https://www.ncei.noaa.gov/products/weather-balloon/integrated-global-radiosonde-archive). Additional model datasets include the North American Multi-Model Ensemble (NMME) (https://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME), SEAS5 from the Copernicus CDS (https://cds.climate.copernicus.eu/datasets/seasonal-monthly-pressure-levels), and CESM2 from the Copernicus CDS CMIP6 Archive (https://cds.climate.copernicus.eu/datasets/projections-cmip6). We utilized a publicly available pre-trained model checkpoint from the ACE2 model (https://huggingface.co/allenai/ACE2-ERA5).

Code availability

The code implementing GAP, including inference scripts and model configurations, is available on Zenodo at https://doi.org/10.5281/zenodo.19112552. Additional analysis scripts and data files used in this study are also included in the repository.

Received: 3 July 2026; Accepted: 10 October 2026;
Published online: 02 November 2026

References

1. Bi, K. et al. Accurate medium-range global weather forecasting with 3d neural networks. *Nature* **619**, 533–538 (2023).

2. Lam, R. et al. Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).
3. Kochkov, D. et al. Neural general circulation models for weather and climate. *Nature*, **632**, 1060–1066 (2024).
4. Price, I. et al. Probabilistic weather forecasting with machine learning. *Nature* **637**, 84–90 (2025).
5. Bodnar, C. et al. A foundation model for the Earth system. *Nature* **641**, 1180–1187 (2025).
6. Sun, X. et al. A data-to-forecast machine learning system for global weather. *Nat. Commun.* **16**, 6658 (2025).
7. Allen, A. et al. End-to-end data-driven weather prediction. *Nature* **641**, 1172–1179 (2025).
8. Chen, L. et al. A machine learning model that outperforms conventional global subseasonal forecast models. *Nat. Commun.* **15**, 6425 (2024).
9. Cresswell-Clay, N. et al. A deep learning Earth system model for efficient simulation of the observed climate. AGU Adv. **6**, e2025AV001706 (2025).
10. Bauer, P., Thorpe, A. & Brunet, G. The quiet revolution of numerical weather prediction. *Nature* **525**, 47–55 (2015).
11. Evensen, G., Vossepoel, F. C. & van Leeuwen, P. J. Data Assimilation Fundamentals: A Unified Formulation of the State and Parameter Estimation Problem (Springer, 2022).
12. Palmer, T., Doblas-Reyes, F., Hagedorn, R. & Weisheimer, A. Probabilistic prediction of climate using multi-model ensembles: from basics to applications. *Philos. Trans. R. Soc. B Biol. Sci.* **360**, 1991–1998 (2005).
13. 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).
14. Manabe, S. & Wetherald, R. T. Thermal equilibrium of the atmosphere with a given distribution of relative humidity. *J. Atmos. Sci.* **24**, 241–259 (1967).
15. von Storch, H., Bruns, T., Fischer-Bruns, I. & Hasselmann, K. Principal oscillation pattern analysis of the 30-to 60-day oscillation in general circulation model equatorial troposphere. *J. Geophys. Res. Atmos.* **93**, 11022–11036 (1988).
16. Hakim, G. J. & Masanam, S. Dynamical tests of a deep-learning weather prediction model. *Artif. Intell. Earth Syst.* **3**, (3) (2024).
17. Bracco, A. et al. Machine learning for the physics of climate. *Nat. Rev. Phys.* **7**, 6–20 (2025).
18. Rasp, S. et al. Weatherbench 2: a benchmark for the next generation of data-driven global weather models. *J. Adv. Model. Earth Syst.* **16**, e2023MS004019 (2024).
19. Luo, J.-J. et al. AI for atmosphere–ocean sciences: advancements, challenges and ways forward. *Natl. Sci. Rev.* **13**, nwag063 (2026).
20. Cheng, S. et al. Machine learning with data assimilation and uncertainty quantification for dynamical systems: a review. *IEEE/CAA J. Autom. Sin.* **10**, 1361–1387 (2023).
21. Bach, E., Baptista, R., Sanz-Alonso, D. & Stuart, A. Machine learning for inverse problems and data assimilation. Preprint at https://arxiv.org/abs/2410.10523 (2024).
22. Bao, F., Zhang, Z. & Zhang, G. A score-based filter for nonlinear data assimilation. J. Comput. Phys. **514**, 113207 (2024).
23. Xiao, Y. et al. Towards a self-contained data-driven global weather forecasting framework. In Proc. 41st International Conference on Machine Learning Vol. **235**, 54255–54275 (PMLR, 2024).
24. Xiao, Y., Jia, Q., Chen, K., Bai, L. & Xue, W. VAE-Var: variational autoencoder-enhanced variational methods for data assimilation in meteorology. In Proc. 13th International Conference on Learning Representations (OpenReview.net, 2025).
25. Watt-Meyer, O. et al. ACE2: accurately learning subseasonal to decadal atmospheric variability and forced responses. npj Clim. Atmos. Sci. **8**, 205 (2025).
26. Sun, Y. Q. et al. Can AI weather models predict out-of-distribution gray swan tropical cyclones? *Proc. Natl. Acad. Sci. USA* **122**, e2420914122 (2025).
27. Palmer, T. & Stevens, B. The scientific challenge of understanding and estimating climate change. *Proc. Natl. Acad. Sci. USA* **116**, 24390–24395 (2019).
28. Ghil, M. Hilbert problems for the climate sciences in the 21st century–20 years later. *Nonlinear Process. Geophys.* **27**, 429–451 (2020).
29. Frigg, R., Bradley, S., Du, H. & Smith, L. A. Laplace's demon and the adventures of his apprentices. *Philos. Sci.* **81**, 31–59 (2014).
30. Hasselmann, K. Stochastic climate models part I. Theory. *Tellus* **28**, 473–485 (1976).
31. Lucarini, V. & Chekroun, M. D. Theoretical tools for understanding the climate crisis from Hasselmann's program and beyond. *Nat. Rev. Phys.* **5**, 744–765 (2023).
32. Murphy, C., Thibeault, V., Allard, A. & Desrosiers, P. Duality between predictability and reconstructability in complex systems. *Nat. Commun.* **15**, 4478 (2024).
33. Zhang, F. et al. What is the predictability limit of midlatitude weather? *J. Atmos. Sci.* **76**, 1077–1091 (2019).
34. Ghil, M. & Lucarini, V. The physics of climate variability and climate change. *Rev. Mod. Phys.* **92**, 035002 (2020).
35. Palmer, T. Stochastic weather and climate models. *Nat. Rev. Phys.* **1**, 463–471 (2019).
36. Buizza, R. et al. *The Development and Evaluation Process Followed at ECMWF to Upgrade the Integrated Forecasting System (IFS)* (Tech. Rep. Technical Report 829, ECMWF, 2018).
37. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. & Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. International Conference on Machine Learning* 2256–2265 (PMLR, 2015).
38. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020).
39. Song, Y. et al. Score-based generative modeling through stochastic differential equations. In Proc. 9th International Conference on Learning Representations (OpenReview.net, 2021).
40. Huang, L., Gianinazzi, L., Yu, Y., Dueben, P. D. & Hoefler, T. DiffDA: a diffusion model for weather-scale data assimilation. In Proc. 41st International Conference on Machine Learning Vol. **235**, 19798–19815 (PMLR, 2024).
41. Rozet, F. & Louppe, G. Score-based data assimilation. *Adv. Neural Inf. Process. Syst.* **36**, 40521–40541 (2023).
42. Manshausen, P. et al. Generative data assimilation of sparse weather station observations at kilometer scales. J. Adv. Model. Earth Syst. **17**, e2024MS004505 (2025).
43. Hess, P., Aich, M., Pan, B. & Boers, N. Fast, scale-adaptive and uncertainty-aware downscaling of Earth system model fields with generative machine learning. Nat. Mach. Intell. **7**, 363–373 (2025).
44. Brenowitz, N. D. et al. Climate in a bottle: towards a generative foundation model for the kilometer-scale global atmosphere. Preprint at https://arxiv.org/abs/2505.06474 (2025).
45. Randall, D. A. & Emanuel, K. The weather–climate schism. *Bull. Am. Meteorol. Soc.* **105**, E300–E305 (2024).
46. Lugmayr, A. et al. RePaint: inpainting using denoising diffusion probabilistic models. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition 11461–11471 (IEEE, 2022).
47. Chung, H., Kim, J., McCann, M. T., Klasky, M. L. & Ye, J. C. Diffusion posterior sampling for general noisy inverse problems. In Proc. 11th International Conference on Learning Representations (OpenReview.net, 2023).
48. Zhang, G. et al. Towards coherent image inpainting using denoising diffusion implicit models. In Proc. 40th International Conference on Machine Learning Vol. **202**, 41164–41193 (PMLR, 2023).

49. Meng, C. et al. SDEdit: guided image synthesis and editing with stochastic differential equations. In Proc. 10th International Conference on Learning Representations (OpenReview.net, 2022).
50. Chao, J. et al. Learning to infer weather states using partial observations. J. Geophys. Res. Mach. Learn. Comput. **2**, e2024JH000260 (2025).
51. Nai, C. et al. Boosting weather forecast via generative superensemble. npj Clim. Atmos. Sci. **8**, 377 (2025).
52. Lee, J. et al. Systematic and objective evaluation of Earth system models: Pcmdi metrics package (pmp) version 3. *Geosci. Model Dev.* **17**, 3919–3948 (2024).
53. Hersbach, H. et al. The era5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
54. Danabasoglu, G. et al. The Community Earth System Model version 2 (cesm2). *J. Adv. Model. Earth Syst.* **12**, e2019MS001916 (2020).
55. Arnold Jr, C. P. & Dey, C. H. Observing-systems simulation experiments: past, present, and future. *Bull. Am. Meteorol. Soc.* **67**, 687–695 (1986).
56. Johnson, S. J. et al. Seas5: the new ECMWF seasonal forecast system. *Geosci. Model Dev.* **12**, 1087–1117 (2019).
57. Lang, S. et al. AIFS-CRPS: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. *npj Artif. Intell.* **2**, 18 (2026).
58. Chen, C., Chen, N., Zhang, Y. & Wu, J.-L. Cgkn: a deep learning framework for modeling complex dynamical systems and efficient data assimilation. *J. Comput. Phys.* **532**, 113950 (2025).
59. Li, Y. et al. Fuxi-en4dvar: An assimilation system based on machine learning weather forecasting model ensuring physical constraints. *Geophys. Res. Lett.* **51**, e2024GL111136 (2024).
60. Vandal, T. J., Duffy, K., Nachmany, Y. & McDuff, D. Global atmospheric data assimilation with multi-modal masked autoencoders. In *Proc. 2025 IEEE International Conference on Data Mining Workshops (ICDMW)*, 863–872 (IEEE, 2025).
61. Alet, F. et al. Skillful joint probabilistic weather forecasting from marginals. Preprint at https://arxiv.org/abs/2506.10772 (2025).
62. Schneider, T., Lan, S., Stuart, A. & Teixeira, J. Earth system modeling 2.0: a blueprint for models that learn from observations and targeted high-resolution simulations. *Geophys. Res. Lett.* **44**, 12–396 (2017).
63. Daras, G. et al. A survey on diffusion models for inverse problems. Preprint at https://arxiv.org/abs/2410.00083 (2024).
64. Gilpin, W. Generative learning for nonlinear dynamics. Nat. Rev. Phys. **6**, 194–206 (2024).
65. Song, J., Meng, C. & Ermon, S. Denoising diffusion implicit models. In Proc. 9th International Conference on Learning Representations (OpenReview.net, 2021).

Acknowledgements

We thank Wenyu Zhou, Peishi Jiang, Jian Lu from Pacific Northwest National Laboratory, Gang Chen from University of California, Los Angeles, Yi Deng from Georgia Institute of Technology, Ming Pan from University of California, San Diego, Y. Q. Sun from University of Chicago, Peili Wu from Met Office, Yu Huang, Philipp Hess, and Michael Aich from Technical University of Munich, Juanjuan Liu, Bin Wang, Li Dong, Guiwan Chen, and Will Yao from Institute of Atmospheric Physics, Chinese Academy of Sciences, Yuling Jiao, Cheng Yuan from Wuhan University, Qingyun Duan, Wentao Li from Hohai University, Wei Gong from Beijing Normal University, Wenbin Chen, Yutao Fan, Yang Liu from Nanjing University, Sencan Sun from Tsinghua University for their insightful discussions and valuable feedback on our work. We also thank the European Centre for Medium-Range Weather Forecasts (ECMWF) for sharing the reanalysis dataset and Weatherbench2 for giving access to reanalysis and model data.

Author contributions

S.Y., C.N., N.B., H.Y. and B.P. conceived and designed the study. S.Y. wrote the main manuscript text and prepared all the figures. C.N., X.Liu, W.L., J.C. and J.Wang prepared the model. L.Wang, Lu L., X.Li, X.C., B.L., Xin L. and Z.X. contributed to the interpretation of results and reviewed the manuscript. N.B. and B.P. edited the manuscript. N.B., H.Y. and B.P. supervised the project. All authors reviewed and approved the final manuscript.

Funding

This research is supported by the National Natural Science Foundation of China (U2342218; H.Y.) and the National Key Research and Development Program of China (Grant Numbers 2023YFC3007700 and 2023YFC3007705). N.B. acknowledges funding from the Volkswagen Foundation and the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101137601 (ClimTip). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

Author Niklas Boers is Guest Editor of npj Climate and Atmospheric Science. Niklas Boers was not involved in the journal's review of, or decisions related to, this manuscript. The other authors do not have a competing interest.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41612-026-01585-4.

Correspondence and requests for materials should be addressed to Niklas Boers, Huiling Yuan or Baoxiang Pan.

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-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/.

© The Author(s) 2026