Data-driven weather forecasts: assimilation, reanalysis and AI models

Atmosphere and Ocean Dynamics, Lecture 10: notes

Pedro S. Peixoto

December 3, 2026

Abstract

The forecast chain, from observations to the forecast, and the data-driven models. These notes discuss data assimilation and 4D-Var, an inverse problem with the model and its adjoint inside; reanalyses and ERA5; the data-driven models trained on ERA5, which learn one 6-hour step and forecast by composing it, and GraphCast's graph network on an icosahedral multi-mesh; what they do well and what they miss (extremes, conservation); end-to-end and hybrid models; and a small AI emulator of the Lorenz-96 system.

Table of contents

1 From 1950 to today: the forecast chain 2

2 Observations: the global observing system 3

3 Data assimilation: an inverse problem 3

4 4D-Var: a minimization problem 4

5 A reanalysis: decades of the past with one model 5

6 ERA5 as training data 6

7 A new revolution: data-driven forecasts 6

8 Seven models in three years 7

9 What is learned, and from what 7

10 Five families of learned models 8

11 Inside a learned model: GraphCast 9

12 One box replaced 11

13 Example: how can an AI model take a 6-hour step? 11

14 What the example shows 12

15 Do they work? Yes, mostly 13

16 What AI models cannot do, yet 14
17 The newest: learning the analysis too 15
18 Where it is going: hybrids, and the world 15
19 A tiny AI weather model 15
20 What the notebook showed 16
21 Key points 17
22 The course in one chain 17
Further reading 18
References 18

1 From 1950 to today: the forecast chain

The first computer forecast, in 1950, integrated one equation (barotropic vorticity) for one layer of air over North America, on a grid of 736 km. A global forecast today solves the full equations for the whole atmosphere, coupled to models of the ocean, the land surface and the sea ice, on grids of about 10 km with about 100 levels, and repeats it about fifty times from slightly different initial states: an **ensemble**. The chain has five links (Figure 1): **observations**; **data assimilation**, which turns them into an **initial state** (Bjerknes' "diagnosis"); the **model**, a dynamical core for the resolved motions plus *parameterizations* of the unresolved processes (radiation, clouds, turbulence, convection); and the **forecast** (Bjerknes' "prognosis") (Bjerknes 1904; Kalnay 2003, Ch. 1).

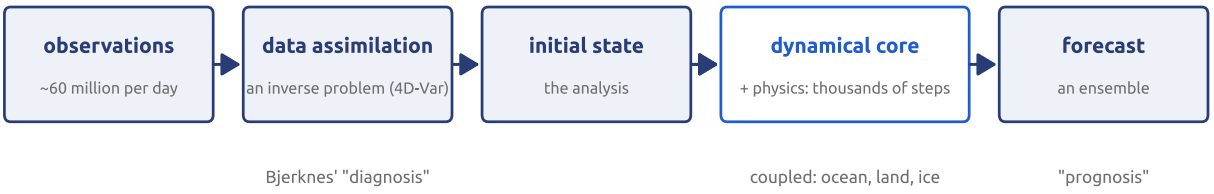


Figure 1: The chain of a physics-based forecast (schematic): observations (about 60 million a day), data assimilation (an inverse problem, 4D-Var), the initial state (Bjerknes' "diagnosis"), the model (a dynamical core plus physics, coupled to ocean, land and ice, thousands of time steps) and the forecast, an ensemble ("prognosis").

The skill of medium-range forecasts has grown by about one day per decade: a six-day forecast today is about as good as a five-day forecast ten years earlier (Bauer et al. 2015). The gains came from all the links at once: better models and higher resolution, far more observations (satellites), better data assimilation, and ensembles. Bauer, Thorpe and Brunet call it a "quiet revolution": Bjerknes' programme of 1904, observe, diagnose, integrate the laws, made industrial. Where does the initial state come from?

2 Observations: the global observing system

Every day ECMWF receives about 800 million observations, and about 60 million of them pass quality control and are available to its forecasting system (Modigliani et al. 2026) (Figure 2). Most come from about 90 satellite instruments. Satellites cannot measure temperature or humidity directly: microwave and infrared sounders measure radiances, from which temperature and humidity are inferred (spectroscopy); radio occultation infers them from the bending of GNSS signals through the atmosphere. Winds come from tracking clouds and water-vapour features and, near the ocean surface, from scatterometers; forecasts are still short of wind information. The in situ observations (surface stations, radiosondes, aircraft, ships, buoys, Argo floats) are few by comparison but direct and local: they calibrate and validate the satellite instruments, expose their biases and anchor the analysis. None of these is a model state: they are irregular in space and time, mostly indirect, and their errors differ from one instrument to the next.

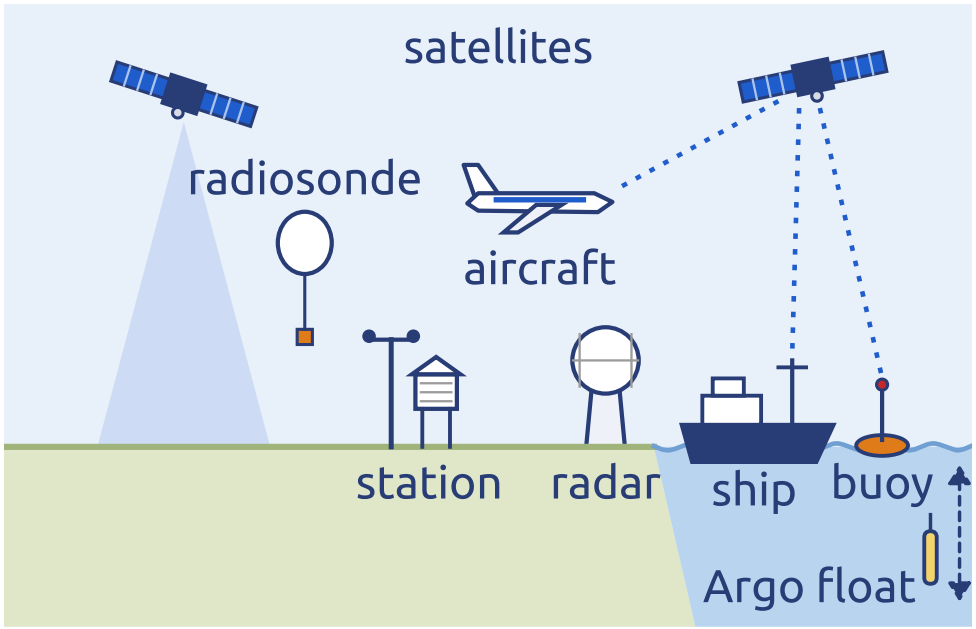


Figure 2: The global observing system (schematic, not to scale; own drawing after the diagram of the World Meteorological Organization, Global Observing System): polar-orbiting and geostationary satellites (the latter also relay the data of ships and buoys), aircraft, radiosondes, surface stations, weather radars, ships, drifting buoys and Argo floats.

3 Data assimilation: an inverse problem

A forecast needs the state of the atmosphere at every grid point, about 10^9 numbers; what we have are the observations, irregular in space and time and mostly *indirect*: a satellite measures radiances, not winds or temperatures. Simply interpolating them onto the grid would leave most of the state undetermined and the rest unbalanced. Data assimilation finds instead the state most consistent with both the observations and a prior estimate, the **background**, which is the previous short-range forecast (Kalnay 2003, Ch. 5).

In **four-dimensional variational assimilation** (4D-Var) the analysis is the model trajectory over an assimilation window (6 to 12 hours) that best fits both the background and the observations of the window. The model is therefore inside the assimilation, and with it its adjoint; operational centres now combine it with ensembles of short forecasts to estimate the background errors

(hybrid assimilation). The much improved analyses of the last decades account for a large part of the gain in forecast skill (Kalnay 2003, Sect. 5.6; Bauer et al. 2015). Figure 3 shows the idea on a toy problem with two unknowns.

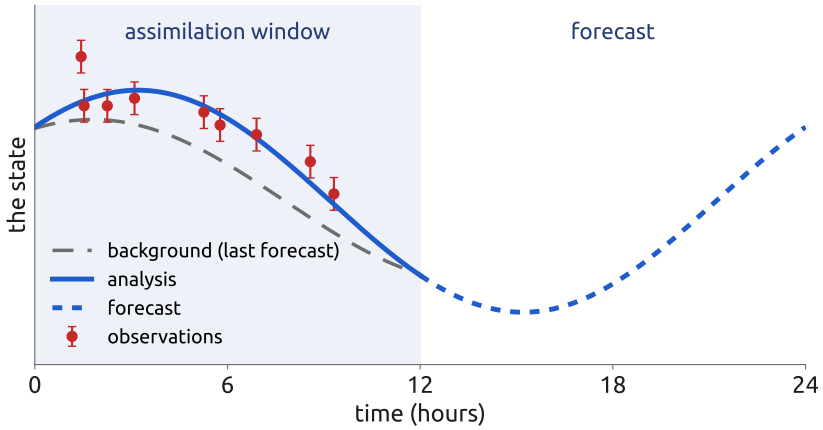


Figure 3: 4D-Var in one picture (schematic). A toy "atmosphere", one oscillation per day, over a 12-hour assimilation window: the background (the last forecast, dashed) has the wrong amplitude and phase; the observations (red, with error bars) are the truth plus noise; the analysis (solid) is the model trajectory whose initial state minimizes the misfit to both, and the forecast continues it after the window. A real 4D-Var controls about 10^9 numbers, not two.

4 4D-Var: a minimization problem

The unknown, or *control variable*, is the initial state $x(t_0)$ of the window. 4D-Var minimizes

$$J[x(t_0)] = \tfrac{1}{2}\,[x(t_0)-x^b]^{\mathsf{T}}\mathbf{B}_0^{-1}[x(t_0)-x^b]+\tfrac{1}{2}\sum_{i=0}^{N}[H(x_i)-y_i^o]^{\mathsf{T}}\mathbf{R}_i^{-1}[H(x_i)-y_i^o], \qquad x_i = M(t_0,t_i)[x(t_0)],$$

in the notation of Kalnay (2003, Sect. 5.6.3, eq. 5.6.14), where $t_0,\dots,t_N$ are the observation times in the window (the index i counts only these times), $x_i = x(t_i)$ the model state at t_i, x^b the background, y_i^o the vector of all observations made at t_i, H the observation operator, which maps a model state to what the instruments measure (radiances, for instance), $M(t_0,t_i)$ the forecast model run from t_0 to t_i (one model, run for longer for each later t_i; Kalnay writes M_i), and $\mathbf{B}_0$ and $\mathbf{R}_i$ the error covariances of the background and of the observations. The first term, J_b, keeps the analysis close to the last forecast where there are no observations: without it the problem is underdetermined, since there are far fewer observations than unknowns. The second, J_o, pulls it towards the observations, each weighted by the inverse of its error covariance.

The minimization is iterative (conjugate gradient or quasi-Newton) and needs the gradient of J with respect to the 10^9 components of $x(t_0)$. With $\mathbf{H}_i$ the Jacobian of H at x_i, $L(t_0,t_i)$ the model linearized about the forward run (the *tangent linear model*), and $L(t_i,t_0)^{\mathsf{T}}$ its adjoint, which carries a vector back from t_i to t_0, Kalnay's eqs. (5.6.17) and (5.6.19) give

$$\frac{\partial J}{\partial x(t_0)} = \mathbf{B}_0^{-1}[x(t_0)-x^b] + \sum_{i=0}^{N} L(t_i,t_0)^{\mathsf{T}}\,\mathbf{H}_i^{\mathsf{T}}\,\mathbf{R}_i^{-1}[H(x_i)-y_i^o].$$

Every iteration runs the model forward through the window, computes the weighted misfits $\mathbf{H}_i^{\mathsf{T}}\mathbf{R}_i^{-1}[H(x_i)-y_i^o]$ at the observation times, and carries them back to t_0 with the adjoint model

L^{T}, which runs backwards in time and picks up each misfit on the way (Figure 4). The whole gradient costs a few model runs, whatever the number of unknowns. Because the analysis is a model trajectory (a **strong constraint**), it carries the model's balances, and the forecast does not start with spurious gravity waves; the price is that the model is taken as perfect, and that $\mathbf{B}_0$ must be known. In practice 4D-Var is **incremental**: the inner minimization uses a linearized model at lower resolution, and outer iterations update the trajectory at full resolution (Kalnay 2003, Sect. 5.6.3, pp. 181–184). A reanalysis needs decades of such analyses: with which model?

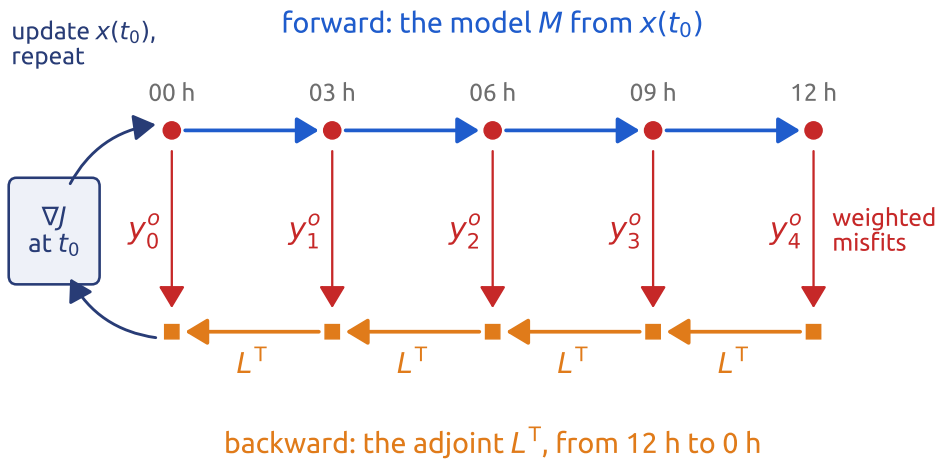


Figure 4: How 4D-Var computes the gradient (schematic, own drawing after Kalnay (2003), Fig. 5.6.2): a 12-hour window with observations every 3 hours. The model runs forward from the current guess of $x(t_0)$ (blue); at each observation time the weighted misfit is computed (red); the adjoint L^{T} runs backwards from 12 h to 0 h (orange) and delivers the gradient at t_0, which the minimizer uses to update $x(t_0)$.

5 A reanalysis: decades of the past with one model

An **analysis** is the initial state of one forecast. The operational analyses of the past cannot simply be stacked into a record of the atmosphere: each was made with the model and the assimilation system of its day, and their improvements would look like climate change. A **reanalysis** repeats the analyses of the past, every day of decades, with **one frozen model** and one assimilation system. **ERA5**, produced by ECMWF for the Copernicus Climate Change Service, covers 1940 to today, hourly, at 0.25° (about 31 km) with 137 levels, about 5 PB of data; it was made with 4D-Var and the IFS of 2016 (cycle 41r2) (Hersbach et al. 2020) (Figure 5). Before the satellite era (1979) it rests on far fewer observations. It is the best estimate of the past atmosphere: one model and all the observations, together (Bauer et al. 2015).

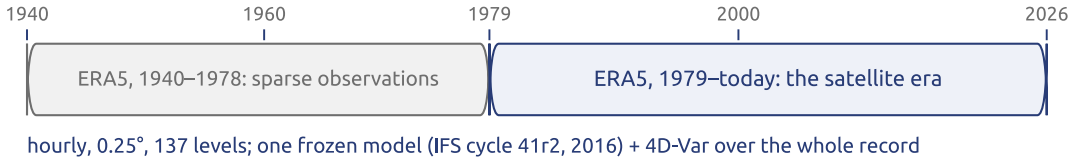


Figure 5: The ERA5 reanalysis on a timeline: hourly global states from 1940 to today, made by one frozen version of ECMWF's model (IFS cycle 41r2, 2016) and 4D-Var; before 1979 far fewer observations, after 1979 a growing constellation of satellites.

ERA5 is free and openly licensed (CC BY 4.0), downloadable from the Copernicus Climate Data Store, on single levels (Copernicus Climate Change Service 2018b) and on pressure levels (Copernicus Climate Change Service 2018a), also from Python with the `cdsapi` package. This public good made data-driven forecasting possible.

6 ERA5 as training data

Four features of ERA5 matter for what a model learned from it can be (Figure 6). It is **circular**: made by the physics model that the learned models are meant to replace, a 2016 version of the IFS with 4D-Var, so "AI beats IFS" partly means "AI matches a 2016 IFS". It is **not homogeneous**: the observations before the satellite era are far fewer, and the satellite constellation keeps growing. It is **not available in real time**: an ERA5 analysis uses observations from up to about 9 hours after its time, while an operational analysis must be ready after about 3 hours; a learned model is trained on one kind of initial state and run on another. And the **training data is also the truth**: the AI models are scored against ERA5, the IFS against its own analyses. GraphCast was trained on 1979–2017 and tested on 2018 and later (Lam et al. 2023; Hersbach et al. 2020). A learned model is only as good as the reanalysis it learns from, and that reanalysis is physics.

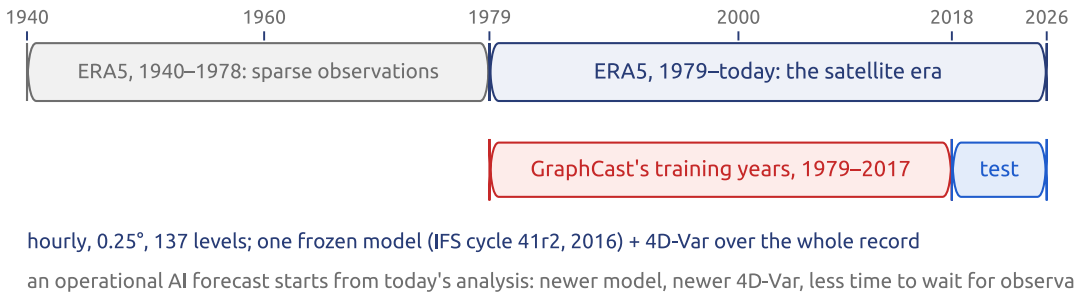


Figure 6: ERA5 (1940 to today) and how GraphCast used it: trained on 1979–2017, tested from 2018; an operational AI forecast then starts from today's operational analysis, made by a newer version of the physics model.

7 A new revolution: data-driven forecasts

For a century, since Richardson, every forecast was computed by discretizing the equations of motion. Since 2022, models that **learn** the evolution of the atmosphere from data have matched, and on many scores beaten, the best physics-based global forecasts (Figure 7). In February 2022 Keisler showed a graph neural network forecasting at 1° (Keisler 2022) and NVIDIA released FourCastNet, a Fourier neural operator (Pathak et al. 2022); Huawei's Pangu-Weather (Bi et al. 2023) and Google DeepMind's GraphCast (Lam et al. 2023) appeared as preprints at the end of 2022 and in journals in 2023; FuXi followed (Chen et al. 2023). On 25 February 2025 ECMWF made its own data-driven model, AIFS, operational (Lang et al. 2024). In three years learned models went from a preprint to an operational forecast. What exactly do they learn?

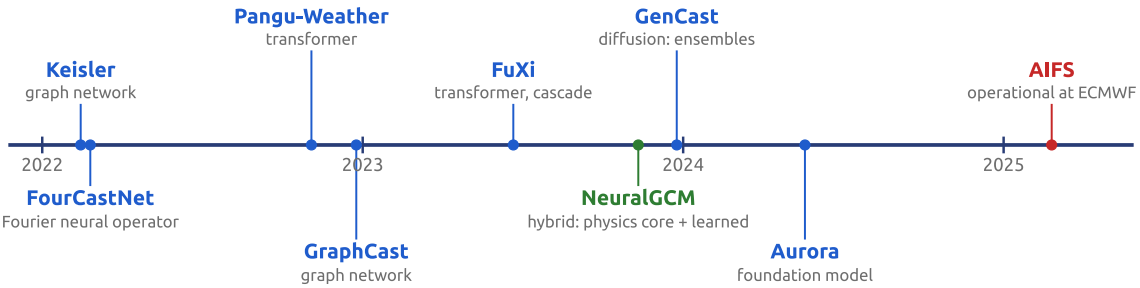


Figure 7: When the main data-driven global models appeared (first public version: the arXiv preprint; the journal papers followed 6 to 12 months later), with the kind of network: graph networks (Keisler, GraphCast), a Fourier neural operator (FourCastNet), transformers (Pangu-Weather, FuXi), a hybrid with a physics core (NeuralGCM, green), a diffusion model for ensembles (GenCast), a foundation model (Aurora), and AIFS, operational at ECMWF since February 2025 (red).

8 Seven models in three years

The slide lists seven models, by the year of their first preprint, and what each brought:

Year	Model	Who	What was new
2022	Keisler GNN (Keisler 2022)	Keisler	a graph neural network on a spherical mesh, encoder–processor–decoder; IFS-like skill at 1°
2022	FourCastNet (Pathak et al. 2022)	NVIDIA, LBNL, Caltech	Fourier neural operator; the first at the IFS's own 0.25°
2022	Pangu-Weather (Bi et al. 2023)	Huawei	3D transformer; the first to claim to beat the IFS deterministic forecast
2022	GraphCast (Lam et al. 2023)	Google DeepMind	a graph network on an icosahedral multi-mesh; better than the IFS on 90% of 1380 targets
2023	GenCast (Price et al. 2025)	Google DeepMind	diffusion: samples possible futures; better than the ECMWF ensemble on 97% of 1320 targets
2023	NeuralGCM (Kochkov et al. 2024)	Google, ECMWF, MIT	hybrid: a differentiable spectral core and learned physics
2025	AIFS (Lang et al. 2024)	ECMWF	graph transformer; the first operational data-driven model at a major centre (open)

Each year brought a new kind of network: graph, Fourier operator, transformer, diffusion, hybrid. AIFS became operational on 25 February 2025 as a deterministic model; its ensemble followed later in 2025.

9 What is learned, and from what

A data-driven model learns a map Φ_θ, with parameters θ, that advances the state by one step of 6 hours:

$$x^{n+1} = \Phi_\theta\left(x^n, x^{n-1}\right), \qquad \theta = \arg\min_\theta \sum_n \left\|\Phi_\theta(x^n, x^{n-1}) - x^{n+1}\right\|^2,$$

where x^n is the whole state at time n, all the variables at all the points (for GraphCast, 227 variables, 5 at the surface and 6 on 37 pressure levels, at about 10^6 points: 2.4×10^8 numbers), and the sum runs over the pairs of consecutive states of the training data (GraphCast uses the two previous states, as written here) (Lam et al. 2023). Φ_θ is not derived from the equations: it is fitted, a least-squares problem in millions of parameters. A 10-day forecast composes the step 40 times, where a physics-based model takes thousands of time steps. In practice the loss is weighted by latitude and by variable, and the training includes rollouts of several steps, so that errors do not grow too fast (Figure 8).

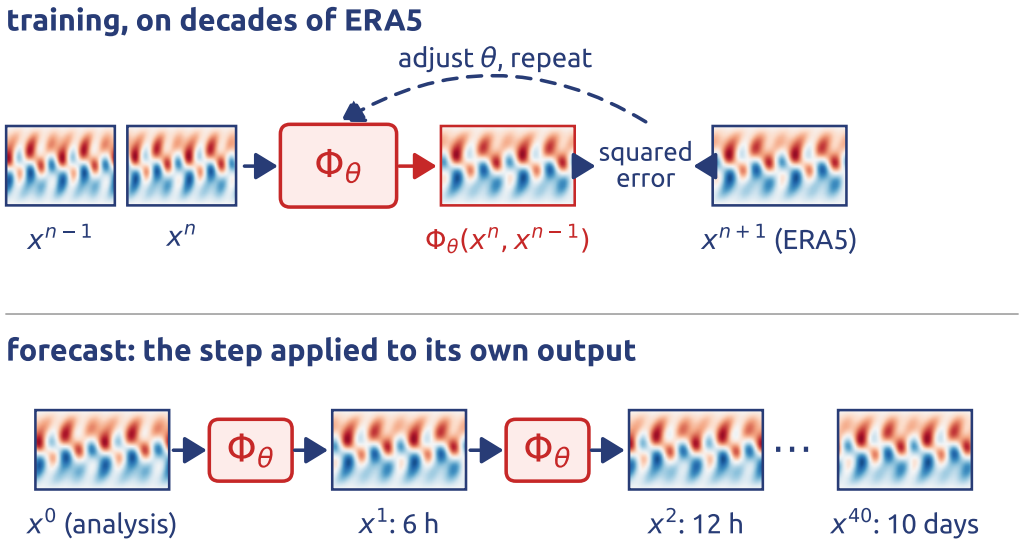


Figure 8: Training and running a learned step (schematic; the small maps are synthetic waves, not data). Top: two consecutive ERA5 states go into the network Φ_θ; its prediction of the state 6 hours later is compared with ERA5 (squared error) and the parameters θ are adjusted, over decades of pairs. Bottom: the trained step applied to its own output, 40 times for a 10-day forecast.

The training data is ERA5 (1940 to today, 0.25°, 137 levels, hourly) (Hersbach et al. 2020), made by the physics model and 4D-Var. One box of the chain is replaced: the model is learned, but the initial state of every forecast still comes from physics.

10 Five families of learned models

The networks differ; the task is the same, a map from the state now to the state 6 hours later (Figure 9):

- **Neural operators** (FourCastNet): each layer transforms the fields to Fourier space, multiplies the modes by learned weights and transforms back, a learned Fourier integral operator (Pathak et al. 2022).
- **Graph networks** (GraphCast): messages passed along the edges of a mesh, a learned stencil (Lam et al. 2023).
- **Transformers** (Pangu-Weather, FuXi, AIFS): attention between patches of the globe, each patch weighting all the others (Bi et al. 2023; Chen et al. 2023); AIFS has a graph-network encoder and decoder and a sliding-window transformer processor (Lang et al. 2024).
- **Diffusion and flow matching** (GenCast; Marchuk, AIRI, 2026): random noise turned step by step into possible states, an ensemble (Price et al. 2025).

- **Hybrids** (NeuralGCM): a discretized dynamical core with a learned closure for the unresolved processes (Kochkov et al. 2024).

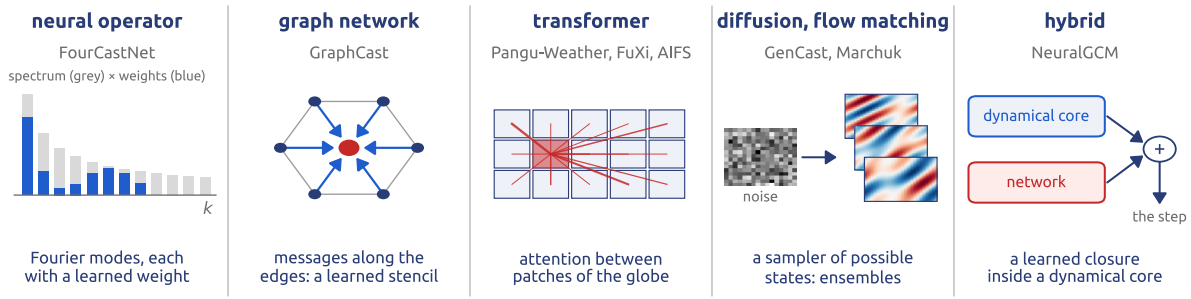


Figure 9: Five families of learned models (schematics). Neural operator: the Fourier spectrum of a field, each mode multiplied by a learned weight. Graph network: messages from the neighbours along the edges of a mesh into one node. Transformer: attention from one patch of the globe to all the others (line width = weight). Diffusion: noise turned step by step into several possible states. Hybrid: a dynamical core and a network adding their tendencies.

What disappears in a learned model is the grid as an approximation of a continuous field, the CFL limit on the time step, and every conservation law; what survives is the state vector, and the initial state, which still comes from 4D-Var. The CFL condition disappears as a limit on the step, but not as a statement about domains of dependence (see the example below).

11 Inside a learned model: GraphCast

GraphCast (Lam et al. 2023) is an *encoder–processor–decoder* graph network, building on Keisler (2022) (Figure 10). (a–c) Its input is the state on the 0.25° grid, 227 variables (5 at the surface, 6 on 37 levels); it predicts the next state, and applied again and again it makes a forecast. (d) The **encoder** maps the grid onto a *multi-mesh*: the vertices of an icosahedron refined six times, 40 962 nodes. (e) The **processor** passes messages along the edges 16 times: each node updates its features from those of its neighbours, like a learned finite-volume flux, but repeated and nonlinear. (f) The **decoder** maps the result back to the grid, as the change of the state in 6 hours. (g) The **multi-mesh** keeps the edges of all seven meshes (the icosahedron and its six refinements) in one graph, from about 7000 km down to about 100 km: near and far at once (Figure 11).

The model has 36.7 million parameters; it was trained on ERA5 from 1979 to 2017, with rollouts of 1 to 12 steps, for about four weeks on 32 TPUs, and it makes a 10-day forecast in under a minute on one TPU, where the IFS takes about an hour on a supercomputer (Lam et al. 2023). Grids are back: a learned operator on an icosahedral mesh, as in finite volumes. But the CFL condition forbids explicit steps of 6 hours on a 28 km grid: how can it take them?

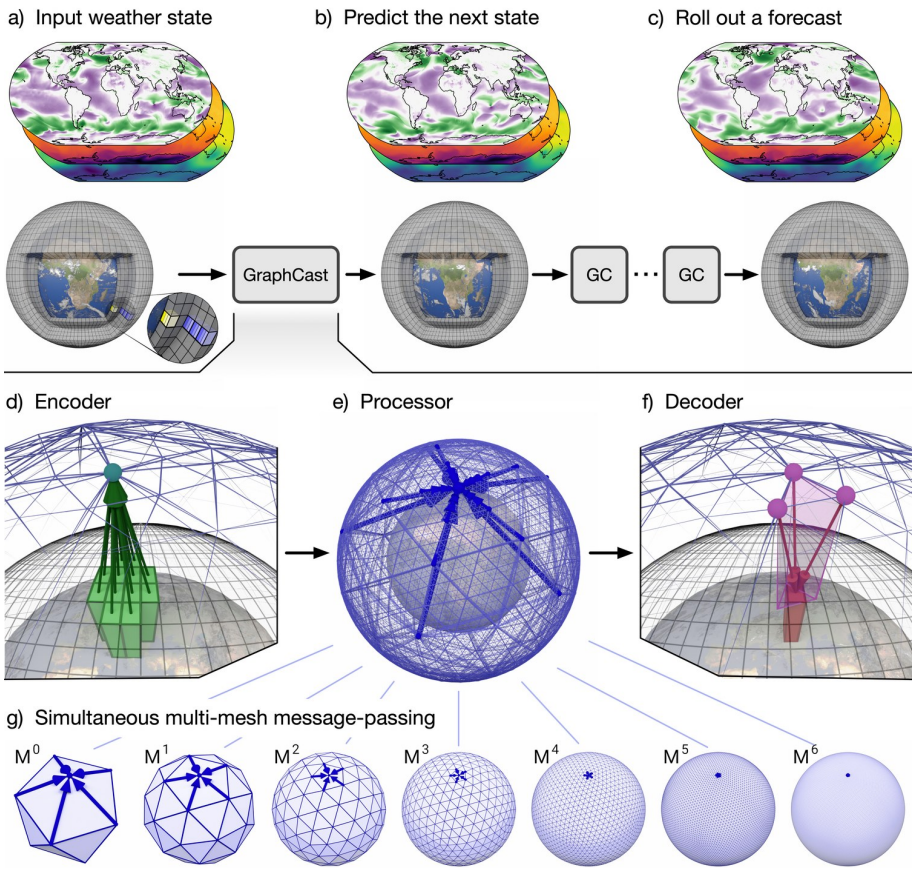


Figure 10: GraphCast, from Lam et al. (2023), Fig. 1 (arXiv version, arxiv.org/abs/2212.12794, CC BY 4.0): (a) the input state on the 0.25° grid, 5 surface and 6 atmospheric variables on 37 levels; (b) GraphCast predicts the next state; (c) applied again and again, a forecast; (d) the encoder maps grid points to the multi-mesh; (e) the processor passes messages on it; (f) the decoder maps back to the grid; (g) the multi-mesh, the edges of seven refinement levels used at once.

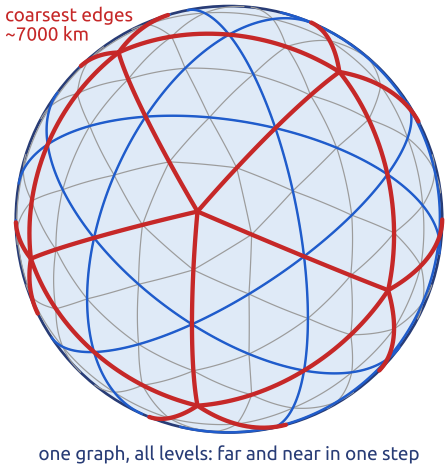


Figure 11: The idea of the multi-mesh (schematic): an icosahedron refined several times (here three levels; GraphCast uses seven), with the edges of all the levels kept in one graph. The coarsest edges (red), the 30 edges of the icosahedron, join points 63.4° of arc, about 7050 km, apart; the finest join neighbouring points.

12 One box replaced

The chain of a data-driven forecast has the same links as that of a physics-based one, with the model box replaced by a learned 6-hour step, applied 40 times for a 10-day forecast, where the physics model takes about 1200 steps of 12 minutes (Figure 12). Four consequences:

- **It still needs physics to run:** the initial state comes from 4D-Var, and 4D-Var needs a model and its adjoint.
- **Its training data needs a physics model:** ERA5 is the IFS plus 4D-Var (Hersbach et al. 2020).
- **Resolution is not the metric:** AIFS, at 28 km, beats the IFS at 9 km on many scores (Lang et al. 2024).
- **The physics–dynamics split is gone:** dynamics and parameterizations are learned together; neither can be inspected or swapped, as a parameterization can in a physics-based model.

The learned step replaces the model; observations, assimilation and reanalysis remain (Bauer et al. 2015).



Figure 12: The forecast chain of a data-driven model (schematic): the same links as in Figure 1, but the model is a learned 6-hour step, applied 40 times for a 10-day forecast. It is trained on a reanalysis (ERA5), itself made by data assimilation with a physics model, and its initial state still comes from data assimilation with a physics model.

13 Example: how can an AI model take a 6-hour step?

The example asks three questions (Figure 13):

- How far does a gravity wave travel in 6 hours at 300 m/s? How many explicit steps would a 28 km grid need for those 6 hours (CFL)?
- The learned model takes one step: how far apart are the two ends of an edge of the icosahedron (63.4° of arc), the coarsest edges of GraphCast's mesh?
- And for a step of 24 hours, with the jet stream (100 m/s) and the gravity waves?

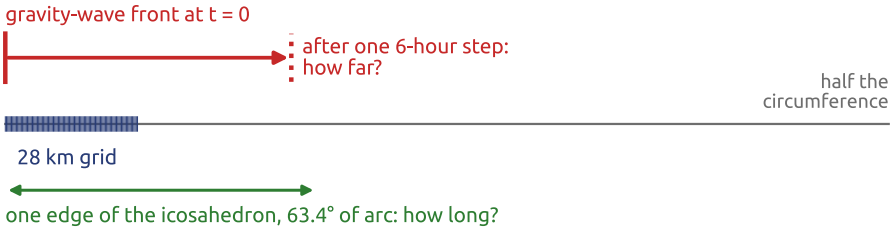


Figure 13: One 6-hour step against the grid and the mesh (schematic, positions to scale): along a great circle, from 0 to 20 000 km (half the Earth's circumference), a gravity-wave front at $t = 0$ and after one 6-hour step; the 28 km grid of a 0.25° model as fine ticks; one edge of the icosahedron as a bracket. The distances are left as questions.

14 What the example shows

Six hours. At 300 m/s a gravity wave travels 300 m/s × 21 600 s = 6480 km. On a 28 km grid the CFL condition with Courant number 1 allows explicit steps of at most 28 km/300 m/s = 93 s: about 230 steps for those 6 hours (Courant et al. 1928).

One edge. An edge of the icosahedron, 63.4° of arc on a sphere of radius 6371 km, is 1.107 × 6371 ≈ 7050 km long: longer than the domain of dependence of a 6-hour step. The multi-mesh keeps the 40 962 nodes of the finest mesh and the edges of all seven levels, about 7000, 3500, 1800, 900, 450, 220 and 110 km long; a node has the edges of the level at which it first appears and of the finer ones, so only the 12 vertices of the icosahedron carry the longest edges, as hubs (Lam et al. 2023). A typical node, with edges of about 110 km, reaches them in a few hops through coarser and coarser nodes (edges of 110, 220, 450, … km), crosses, and comes down: a few of the 16 rounds of message passing. The receptive field of one step covers the domain of dependence: the same idea as a semi-Lagrangian scheme, which puts the domain of dependence inside the stencil.

Twenty-four hours. The jet stream carries information 100 m/s × 86 400 s = 8640 km, about two hops of 7050 km; gravity waves travel 25 900 km, more than half the circumference (20 000 km). A 24-hour step needs a global receptive field; three hops of 7050 km reach anywhere, and GraphCast's processor makes 16 (Figure 14). A learned step is a stencil: it must reach as far as the fastest signal travels in one step.

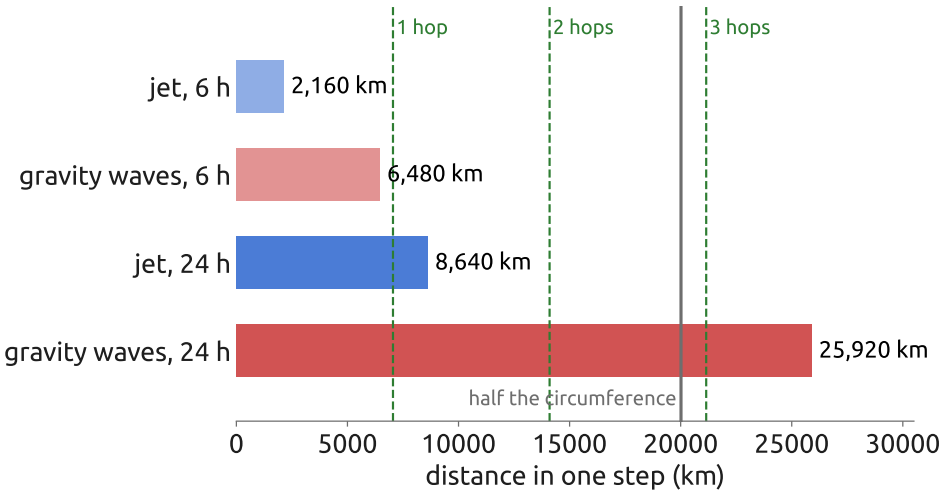


Figure 14: Distances along a great circle: a gravity wave (300 m/s) and the jet stream (100 m/s) in one step of 6 hours and of 24 hours, against the length of one edge of the icosahedron (7050 km; dashed lines at 1, 2 and 3 hops) and half the Earth's circumference (20 000 km). Explicit steps on a 28 km grid with Courant number 1 for gravity waves: 93 s, about 230 for 6 hours.

15 Do they work? Yes, mostly

On the standard scores (root-mean-square errors and anomaly correlations of the main variables at many lead times), GraphCast beat ECMWF's deterministic forecast, HRES, on 90% of 1380 targets (Lam et al. 2023) (Figure 15), and GenCast, a diffusion model, beat ECMWF's ensemble on 97% of 1320 targets (Price et al. 2025). Open benchmarks such as WeatherBench 2 compare the models on the same footing (Rasp et al. 2024), and ECMWF has assessed them in an operational-like setting (Ben Bouallègue et al. 2024). AIFS, at 28 km, beats physics-based models at 9 km on many scores (Moldovan et al. 2026); once trained, it needs about a thousand times less energy per forecast than the physics-based model (ECMWF news, 25 February 2025). The scores carry a caveat: the AI models are scored against ERA5, the physics-based models against their own analyses.

What does not work, yet: the forecasts **blur**, because a model trained on the mean-square error forecasts the average of the possible futures; **extremes** are rare in the data; nothing enforces **conservation laws**; and the models need 4D-Var, a reanalysis and the climate of their training years. The next section takes these one by one.

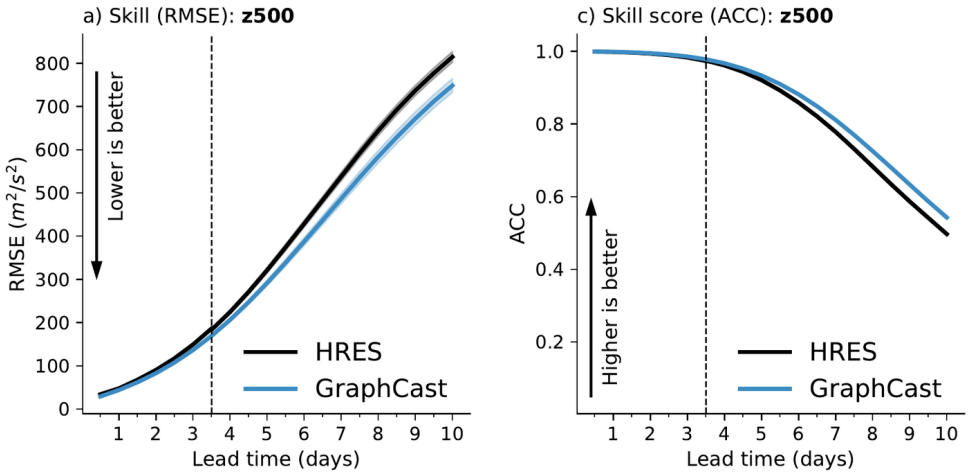


Figure 15: GraphCast against ECMWF's HRES, from Lam et al. (2023), Fig. 2a and 2c (arXiv version, arxiv.org/abs/2212.12794, CC BY 4.0): forecasts of 2018 of the 500 hPa geopotential, root-mean-square error (m^2/s^2, lower is better) and anomaly correlation (ACC, higher is better) against lead time; GraphCast (blue) is scored against ERA5, HRES (black) against its own analysis; shading: 95% confidence intervals; dashed line at 3.5 days, where the HRES initialization times change.

16 What AI models cannot do, yet

1. **Extremes.** Record-breaking heat, cold and wind are forecast better by physics-based models (IFS-HRES) than by the AI models (Zhang et al. 2026): extremes are rare in the training data, and the mean-square error pulls the forecast towards the mean.
2. **Blurring.** The forecast that minimizes the mean-square error is the *mean* of the possible futures, which is smooth: fronts and storm centres are smeared out as the lead time grows (Figure 16). Diffusion models such as GenCast sample sharp possible states instead of averaging them (Price et al. 2025).
3. **No conservation laws.** Nothing in the network enforces the conservation of mass or energy; long runs drift. Hybrids keep a dynamical core for this reason, among others (Kochkov et al. 2024).
4. **Circular data.** ERA5 is a physics product, and it is not available in real time.
5. **A changing climate.** The models are trained on the past (GraphCast on 1979–2017) and used in the climate of today; they learn the climate of their training period.

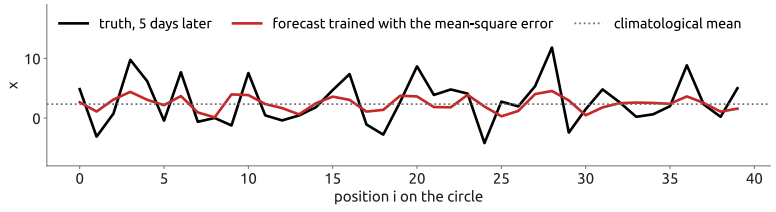


Figure 16: Blurring in a tiny AI model (Lorenz-96, 40 variables on a circle; one step of 0.05 time units is about 6 hours): the truth 5 days ahead (black) and a linear forecast fitted directly for 5 days with the mean-square error (red), pulled towards the climatological mean (dotted): it loses the amplitude of the waves.

17 The newest: learning the analysis too

Every model so far starts from a physics-based analysis: 4D-Var, the model trajectory that best fits the background and the observations (Figure 3). The newest systems learn this box too. **End to end**, they learn the map from the raw observations to the forecast, with no 4D-Var at run time. Aardvark Weather (Cambridge and the Alan Turing Institute) has an encoder from the observations to a gridded state, a processor and decoders to the forecasts (Allen et al. 2025); FuXi-Weather (Fudan) cycles data assimilation and forecast with learned models (Sun et al. 2025). Both are still trained on reanalyses, and still behind the best operational systems on many scores. The last box of the chain is being learned, from the observations straight to the forecast; without a physics model anywhere, who checks the learned atmosphere?

18 Where it is going: hybrids, and the world

Three ways to fill the model box are now in use (Figure 17). **Hybrid** models combine the two approaches: NeuralGCM keeps a differentiable spectral dynamical core and learns only the effect of the unresolved processes, training through the solver (Kochkov et al. 2024): the equations come back. **Foundation models** such as Aurora pretrain one large network on many data sets and fine-tune it to weather, air quality or ocean waves (Bodnar et al. 2025). And **ensembles** come from diffusion models, GenCast and FuXi-ENS, which sample possible futures instead of averaging them (Price et al. 2025; Zhong et al. 2025). The likely future is physics and learning together, not one replacing the other.

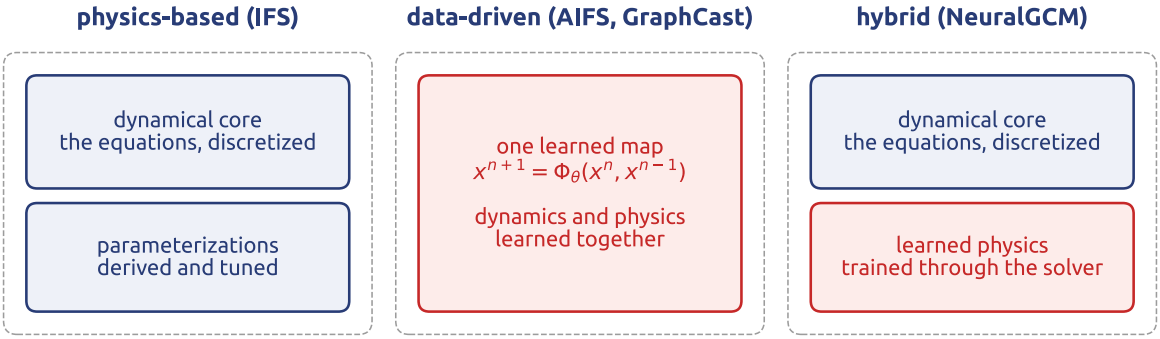


Figure 17: Three ways to fill the model box (schematic): physics-based (a dynamical core plus parameterizations, derived and tuned), data-driven (one learned 6-hour map, dynamics and physics learned together) and hybrid (a differentiable core plus learned physics, trained through the solver).

19 A tiny AI weather model

The whole idea fits on a laptop, in numpy, with the Lorenz-96 model as the "atmosphere" (Lorenz 1996): $K = 40$ variables x_i on a latitude circle (periodic),

$$\frac{dx_i}{dt} = (x_{i+1} - x_{i-2})\, x_{i-1} - x_i + F,$$

with $F = 8$, chaotic (Figure 18). The quadratic term is an advection, $-x_i$ a damping and F a forcing; Lorenz chose the time unit so that 0.05 is about 6 hours of the atmosphere. The

computation takes the steps in turn: the model, integrated with RK4 and steps of 0.01, after a spin-up that drops the transient from rest; a "reanalysis", a long run of 16 000 states 6 hours apart (about 11 "years") for training and a separate run for testing; the 6-hour step learned by least squares, first as a linear map $x^{n+1} = x^n + Wx^n + \mathbf{b}$ (1640 parameters, seeing the whole circle), then by a small network with a local stencil, $x_{i-2}, \ldots, x_{i+2}$, two hidden layers of 32 and the same weights at every point (1281 parameters); forecasts by composing the step 40 times; and a model fitted to the mean-square error 5 days ahead, which blurs. Which forecasts better, the linear model or the small network?

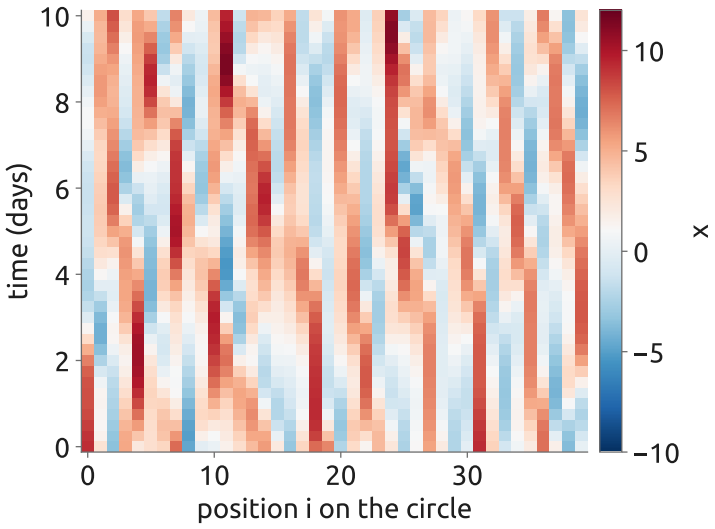


Figure 18: Lorenz-96 ($K = 40$, $F = 8$) over 10 days: x_i against the point i and time, an "atmosphere" of 40 numbers around a latitude circle, with waves that grow, travel and decay irregularly.

20 What the notebook showed

The one-step errors are 0.78 for the linear model and 0.10 for the local network (RMS; the spread of the states is 3.6). From 200 test states with initial errors of 0.05, the RMS errors at days 1 and 5 are 3.3 and 5.6 for persistence, 2.6 and 3.7 for the linear model, 0.40 and 2.2 for the local network, and 0.07 and 0.39 for the true model started from the same wrong state (Figure 19). The local network, with fewer parameters, beats the linear model: it was given the right structure, a local rule that is the same everywhere, like a finite-difference stencil, and its sensitivities to $x_{i-2}, \ldots, x_{i+2}$ equal those of the true 6-hour step. Its stencil must cover where information comes from in one step: a stencil of width 3, which misses x_{i-2}, fails (RMS error 2 at day 1), width 5 works: the CFL condition, learned. The linear model decays to the mean: its error stops at about the spread of the states. Nothing beats the true model from a slightly wrong start for long: its error is the initial error amplified by chaos, and the network's one-step error costs it about five days of predictability. A learned step is a numerical method whose coefficients were fitted instead of derived.

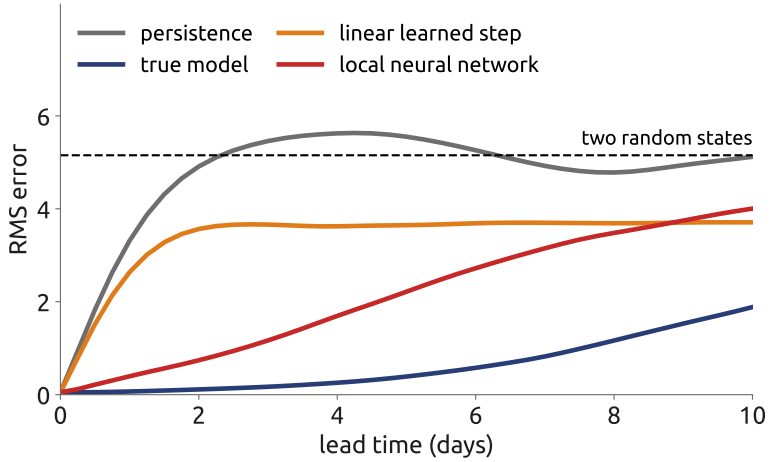


Figure 19: Lorenz-96 (one step of 0.05 is about 6 hours): RMS error against lead time of 10-day forecasts from 200 test states, all started from the same perturbed state (errors of 0.05): persistence, the true model, a linear learned step and a local neural network (stencil of five points, the same weights at every point). The dashed line is the RMS difference of two unrelated states of the model, $\sqrt{2}$ times the spread of its states.

21 Key points

1. **The initial state:** data assimilation, an inverse problem with the model inside (4D-Var).
2. **Reanalysis:** decades of analyses with one frozen model; ERA5 is what AI models learn from.
3. **Learned models:** a 6-hour step fitted to ERA5; it must see as far as the fastest signal goes: **CFL, learned**.
4. **One box replaced:** the observations, the assimilation and the reanalysis remain physics.
5. **Do they work?** Excellent average forecasts; extremes, conservation and a changing climate remain open.
6. **The newest:** learning the analysis too (end to end); hybrids keep the equations.

AI replaces one box of the chain: the observations and the physics around it remain.

22 The course in one chain

Topic	What you can now do
History	tell the story from Halley to the 1950 forecast, and why rotation matters
Laws	derive the equations, reduce them to shallow water, find their waves and balances
Vertical	treat the vertical: buoyancy, internal waves, convection, mountain waves
Grids	discretize them with finite differences and volumes; stability, CFL, conservation
Spectral, FE	spectral methods and finite elements: the cores of IFS and LFRic
Models	follow the models from 1950 to today: grids, the vertical, the tests
AI	read the chain of a modern forecast and what learned models change in it

Bjerknes' two conditions of 1904, a sufficiently accurate knowledge of the state of the atmosphere at an initial time and of the laws by which one state develops from another, organize the whole

course (Bjerknes 1904): the laws, their reduction and their discretization, and the initial state and the forecast chain (Bauer et al. 2015). What remains open is as much mathematics as engineering: dynamical cores that scale to the kilometre on new computers; learned models that conserve what the equations conserve, forecast extremes and stay reliable in a changing climate; and the coupling of both. The problem is old; the tools are yours now.

Further reading

Bauer et al. (2015) on the progress of numerical weather prediction; Kalnay (2003), Chapter 5, on data assimilation and 4D-Var; Hersbach et al. (2020) on ERA5; Lam et al. (2023) and Price et al. (2025) for two learned models in detail, Rasp et al. (2024) for how they are compared, and Zhang et al. (2026) for their extremes.

References

Allen, Anna, Stratis Markou, Will Tebbutt, et al. 2025. "End-to-End Data-Driven Weather Prediction." *Nature* 641 (8065): 1172–79. https://doi.org/10.1038/s41586-025-08897-0.

Bauer, Peter, Alan Thorpe, and Gilbert Brunet. 2015. "The Quiet Revolution of Numerical Weather Prediction." *Nature* 525: 47–55. https://doi.org/10.1038/nature14956.

Ben Bouallègue, Zied, Mariana C. A. Clare, Linus Magnusson, et al. 2024. "The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context." *Bulletin of the American Meteorological Society* 105 (6): E864–83. https://doi.org/10.1175/bams-d-23-0162.1.

Bi, Kaifeng, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. 2023. "Accurate Medium-Range Global Weather Forecasting with 3D Neural Networks." *Nature* 619: 533–38. https://doi.org/10.1038/s41586-023-06185-3.

Bjerknes, Vilhelm. 1904. "Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik." *Meteorologische Zeitschrift* 21: 1–7.

Bodnar, Cristian et al. 2025. "A Foundation Model for the Earth System." *Nature* 641: 1180–87. https://doi.org/10.1038/s41586-025-09005-y.

Chen, Lei, Xiaohui Zhong, Feng Zhang, et al. 2023. "FuXi: A Cascade Machine Learning Forecasting System for 15-Day Global Weather Forecast." *Npj Climate and Atmospheric Science* 6: 190. https://doi.org/10.1038/s41612-023-00512-1.

Copernicus Climate Change Service. 2018a. *ERA5 Hourly Data on Pressure Levels from 1940 to Present*. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.bd0915c6.

Copernicus Climate Change Service. 2018b. *ERA5 Hourly Data on Single Levels from 1940 to Present*. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.adbb2d47.

Courant, Richard, Kurt Friedrichs, and Hans Lewy. 1928. "Über die partiellen Differenzengleichungen der mathematischen Physik." *Mathematische Annalen* 100: 32–74. https://doi.org/10.1007/BF01448839.

Hersbach, Hans et al. 2020. "The ERA5 Global Reanalysis." *Quarterly Journal of the Royal Meteorological Society* 146: 1999–2049. https://doi.org/10.1002/qj.3803.

Kalnay, Eugenia. 2003. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press. https://doi.org/10.1017/CBO9780511802270.

Keisler, Ryan. 2022. *Forecasting Global Weather with Graph Neural Networks*. https://doi.org/10.48550/arXiv.2202.07575.

Kochkov, Dmitrii et al. 2024. "Neural General Circulation Models for Weather and Climate." *Nature* 632: 1060–66. https://doi.org/10.1038/s41586-024-07744-y.

Lam, Remi et al. 2023. "Learning Skillful Medium-Range Global Weather Forecasting." *Science* 382: 1416–21. https://doi.org/10.1126/science.adi2336.

Lang, Simon et al. 2024. *AIFS – ECMWF's Data-Driven Forecasting System*. https://doi.org/10.48550/arXiv.2406.01465.

Lorenz, Edward N. 1996. "Predictability: A Problem Partly Solved." *Proceedings of the ECMWF Seminar on Predictability, 4–8 September 1995* (Reading) 1: 1–18.

Modigliani, Umberto, Irina Sandu, Stephen English, and Tony McNally. 2026. *From Space to Forecast: How ECMWF Turns Satellite Data into Trusted Weather Predictions*. ECMWF news, 3 September 2026. https://www.ecmwf.int/en/about/media-centre/news/2026/satellite-data-weather-forecasts.

Moldovan, Gabriel, Ewan Pinnington, Ana Prieto Nemesio, et al. 2026. "AIFS Single 1.1.0: An Update to ECMWF's Machine-Learned Weather Forecast Model AIFS." *Geoscientific Model Development* 19 (10): 4703–24. https://doi.org/10.5194/gmd-19-4703-2026.

Pathak, Jaideep et al. 2022. *FourCastNet: A Global Data-Driven High-Resolution Weather Model Using Adaptive Fourier Neural Operators*. https://doi.org/10.48550/arXiv.2202.11214.

Price, Ilan et al. 2025. "Probabilistic Weather Forecasting with Machine Learning." *Nature* 637: 84–90. https://doi.org/10.1038/s41586-024-08252-9.

Rasp, Stephan et al. 2024. "WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models." *Journal of Advances in Modeling Earth Systems* 16: e2023MS004019. https://doi.org/10.1029/2023MS004019.

Sun, Xiuyu, Xiaohui Zhong, Xiaoze Xu, et al. 2025. "A Data-to-Forecast Machine Learning System for Global Weather." *Nature Communications* 16 (1). https://doi.org/10.1038/s41467-025-62024-1.

Zhang, Zhongwei, Erich Fischer, Jakob Zscheischler, and Sebastian Engelke. 2026. "Physics-Based Models Outperform AI Weather Forecasts of Record-Breaking Extremes." *Science Advances* 12 (18). https://doi.org/10.1126/sciadv.aec1433.

Zhong, Xiaohui, Lei Chen, Hao Li, et al. 2025. "FuXi-ENS: A Machine Learning Model for Efficient and Accurate Ensemble Weather Prediction." *Science Advances* 11 (44). https://doi.org/10.1126/sciadv.adu2854.