Daily global temperature, 1940–2026

ERA5 from Copernicus Climate Pulse: descriptive analysis, a Bayesian DLM and SARIMA-type models

Hedibert Freitas Lopes

November 2026

Disclaimer: these slides, the R code and the analyses were prepared with AI assistance (Claude, Anthropic) and reviewed by the author.

Insper – Instituto de Ensino e Pesquisa hedibert.org

Motivation

- A *Scientific American* post (Nov 2026), based on preliminary Copernicus data: **Friday, 20 November 2026 was 2.12 °C above the preindustrial average**.
- Their chart overlays the daily global temperature anomaly of every year since 1940.
- In the final data, 20 Nov 2026 is **2.13 °C**.
- Days above 2 °C (vs 1850–1900): 2 in 2023, 3 in 2024, 0 in 2025 and **7 in 2026** (to 22 Nov).

Questions for today

1. Where do these numbers come from, and how reliable are they?
2. What do the daily data look like (1940 onward and 1979 onward)?
3. How fast is the trend changing? Is the seasonal cycle changing?
4. Bayesian DLM vs classical SARIMA-type models.
5. An exercise: when would purely statistical models put us past 2 °C?

The data

Source: Copernicus Climate Pulse

- A wonderful site: https://pulse.climate.copernicus.eu
  Copernicus Climate Change Service (C3S) / ECMWF. Near-real-time daily global air temperature and sea-surface temperature, with free CSV downloads.
- File used: `timeseries_era5_t2m_daily_global.csv`, covering **1 Jan 1940 – 22 Nov 2026** (31,738 days).
- Columns:
  - `t2m`: global mean 2 m air temperature (°C)
  - `clim_1991_2020`: daily climatology over 1991–2020
  - `anom_1991_2020`: anomaly relative to 1991–2020
  - `offset_1850_1900`: offset between 1850–1900 and 1991–2020 (≈0.88–0.90 °C)
  - `anom_1850_1900` = `anom_1991_2020` + `offset_1850_1900`
  - `status`: FINAL or PRELIMINARY
- Dataset: ERA5, doi:10.24381/cds.adbb2d47.

What is ERA5? A reanalysis, not a set of thermometers

- The numbers are **not** measured at a fixed set of locations.
- ERA5 is a **reanalysis**: a weather-forecast model that takes in observations from weather stations, ships, buoys, weather balloons, aircraft and (since the 1970s) satellites.
- **Data assimilation** (4D-Var) merges the observations with the model physics to estimate the full state of the atmosphere **every hour** on a grid of about 31 km (0.25°), with no gaps.
- The daily global value is the average over **every grid cell** of the Earth.
- **2 m or 10 m?** The variable is the **2 m air temperature** (the height of a thermometer screen). The model's lowest level is about 10 m up, and the 2 m value is calculated between that level and the surface. This is probably where the "10 m" on the site comes from.
- Over the ocean it is the *air* temperature just above the water, not the sea-surface temperature. This is one reason ERA5 differs slightly from HadCRUT or NOAA, which mix land air temperature with sea-surface temperature.

How the global average is weighted

- **Area-weighted mean** with weights cos(latitude): every km^2 of the Earth's surface counts the same.
- **Hemispheres:** each counts for exactly half. The poles and Antarctica are included.
- **Land vs ocean:** in proportion to area, about **29% land and 71% ocean**.
- The Northern Hemisphere has much more land (about 39% vs 19% in the Southern), and land heats and cools faster than water. So:
  - the global mean **peaks in July** (northern summer);
  - it swings about **3.9 °C** over the year (see the seasonal-cycle results later).
- **For a statistician:** the grid is complete, but the information behind it is not equally strong. Where observations are sparse, the model does more of the work, and ERA5 publishes a single value *without error bars*.

How reliable is the series over time?

Period	Assessment
1940–1957	**Use with caution.** Few observations, and the war disrupted ship reports. Daily and Southern Hemisphere values lean heavily on the model. This stretch is a later back-extension of ERA5, with lower quality, especially before 1950.
1958–1978	**Reasonable for global and yearly averages.** The International Geophysical Year (1957–58) built up the weather-balloon network. Annual global means agree well with independent datasets (HadCRUT, Berkeley Earth, JRA-55). Daily values are noisier.
1979 onward	**Robust**, including daily values and regional detail. Large amounts of satellite data enter (microwave and infrared temperature sounders), so the model is constrained everywhere.

- Yearly global trends are fine from about 1950. For daily extremes or day-of-year distributions, start in **1979**.
- **Watch for spurious breaks:** a step change around 1979 can come from the satellite data arriving, not from the climate.

Two more caveats

1. **"Preindustrial" is estimated.** ERA5 starts in 1940, so it has no 1850–1900 data. The anomaly vs 1850–1900 adds an offset taken from other datasets (`offset_1850_1900` $\approx$ 0.88–0.90 °C) to the anomaly vs 1991–2020. The "2.12 °C" is therefore partly a convention.
2. **The most recent days are preliminary (ERA5T).** They are published about 5 days behind real time and may be revised when the final release comes out 2–3 months later. That is why Sci Am's preliminary 2.12 °C became 2.13 °C.

In what follows: the variable is the absolute `t2m` (°C), and every analysis is repeated with data starting in **1940** and in **1979**.

Descriptive analysis, 1940–2026

Figure 1: daily global temperature, coloured by decade (1940–2026)

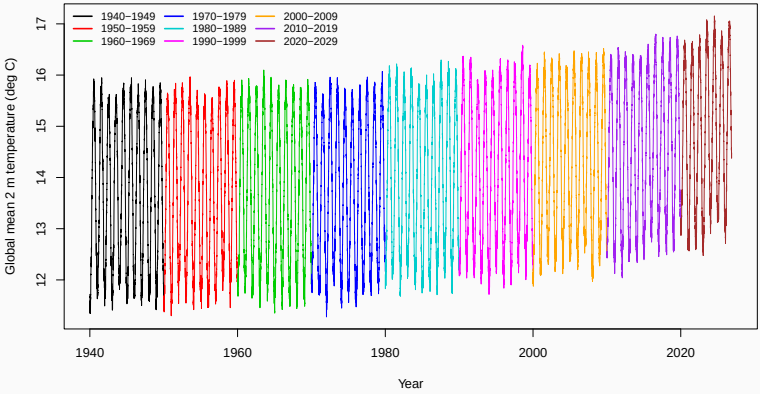
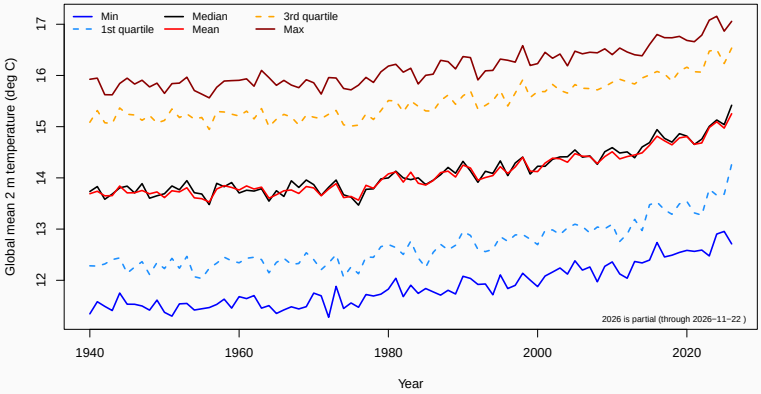
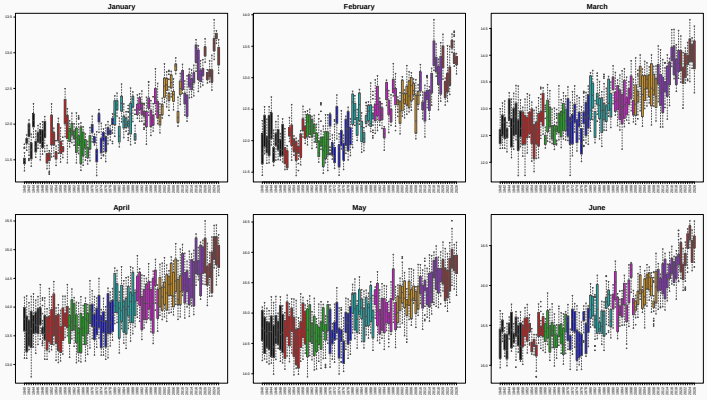


Figure 2: yearly summaries of the daily values (1940–2026)



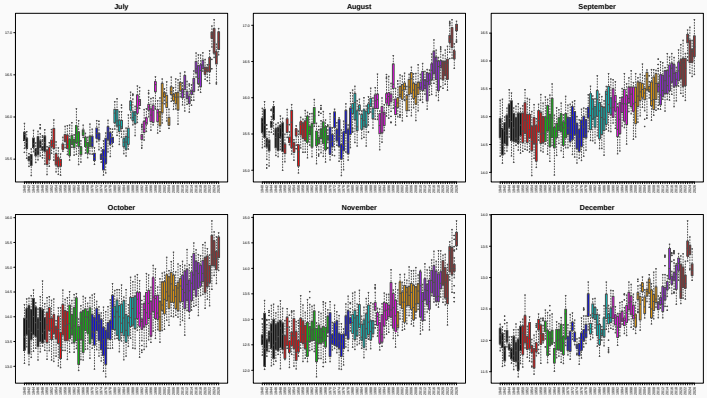
Each year's min, Q1, median, mean, Q3 and max. 2026 runs only to 22 Nov: leaving out the cooler end of the year inflates its min and Q1.

Figure 3a: each month across years, January–June (1940–2026)



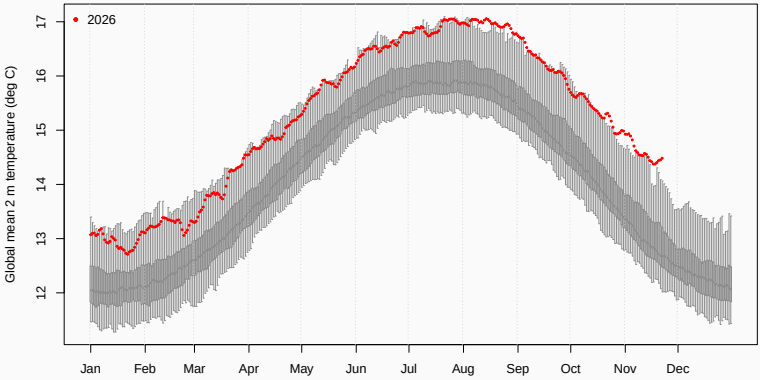
Box colours = decades, as in Figure 1. Each panel has its own y-axis.

Figure 3b: each month across years, July–December (1940–2026)



Box colours = decades, as in Figure 1. Each panel has its own y-axis.

Figure 4: 365 day-of-year boxplots (1940–2026, Feb 29 excluded)



Red points: 2026. They sit at or above the top of the boxes almost every day.

Descriptive analysis, 1979–2026

Figure 1: daily global temperature, coloured by decade (1979–2026)

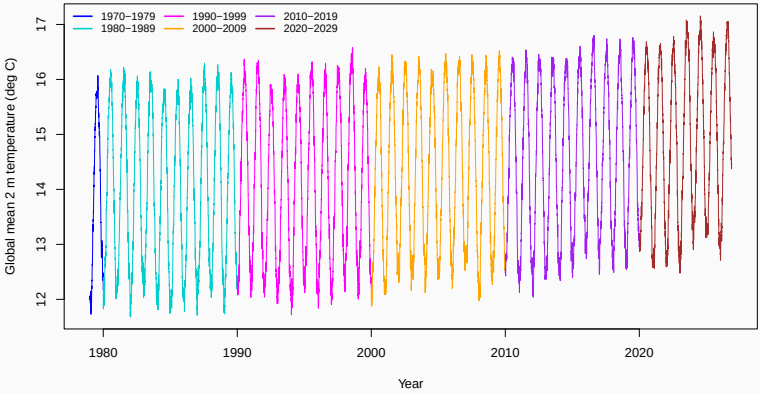


Figure 2: yearly summaries of the daily values (1979–2026)

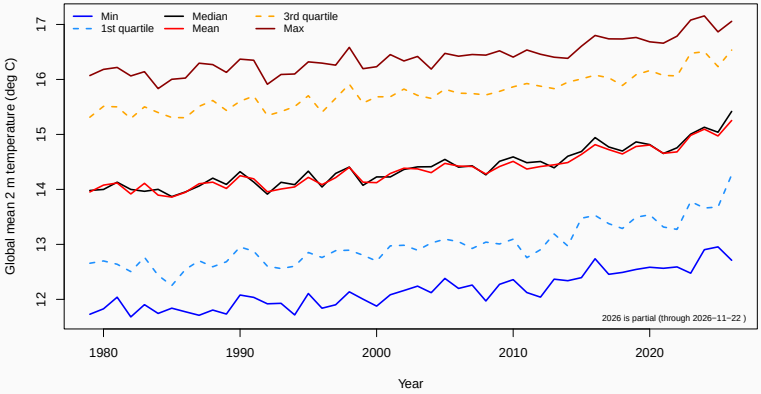
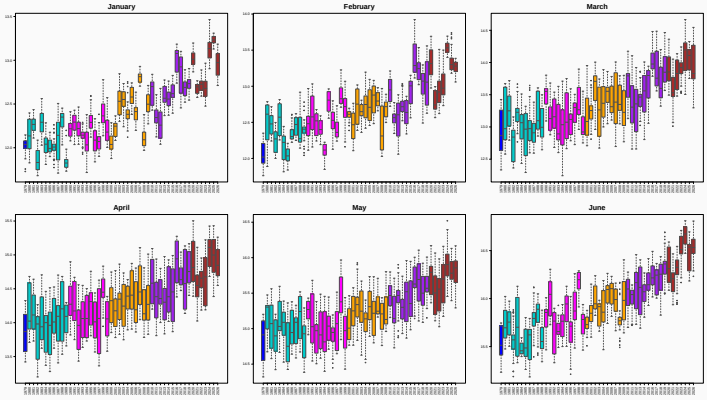
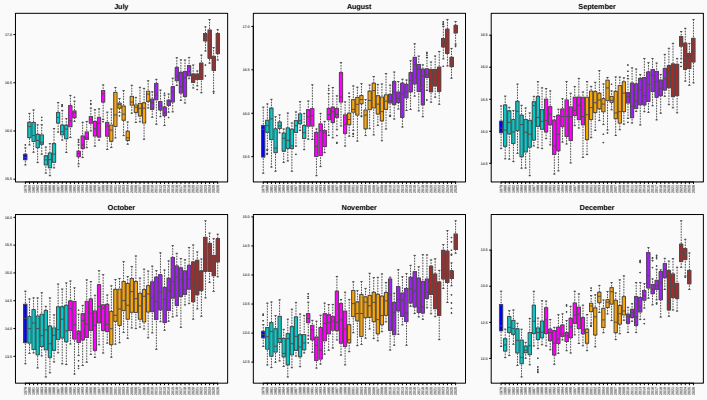


Figure 3a: each month across years, January–June (1979–2026)



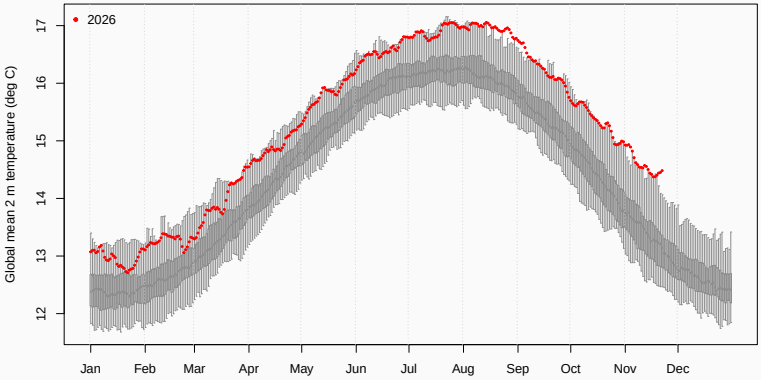
Box colours = decades, as in Figure 1.

Figure 3b: each month across years, July–December (1979–2026)



Box colours = decades, as in Figure 1.

Figure 4: 365 day-of-year boxplots (1979–2026, Feb 29 excluded)



Red points: 2026. The boxes are tighter than in the 1940–2026 version: fewer cold early decades.

A Bayesian dynamic linear model

Model for the daily series (Figure 1)

Observation equation, for day t in month $m(t)$:

$$y_t = \mu_t + \sum_{j=1}^{3} a_{jt} + \gamma_{m(t)} + x_t + v_t, \qquad v_t \sim N(0, V)$$

- **Local linear trend:** $\mu_t = \mu_{t-1} + \beta_{t-1} + w_{\mu t}$, $\beta_t = \beta_{t-1} + w_{\beta t}$.
- **Trigonometric seasonal (West & Harrison form)**, $K = 3$ harmonics, $\omega_j = 2\pi j/365.25$:

$$\begin{pmatrix} a_{jt} \\ b_{jt} \end{pmatrix} = \begin{pmatrix} \cos\omega_j & \sin\omega_j \\ -\sin\omega_j & \cos\omega_j \end{pmatrix} \begin{pmatrix} a_{j,t-1} \\ b_{j,t-1} \end{pmatrix} + w_{jt}.$$

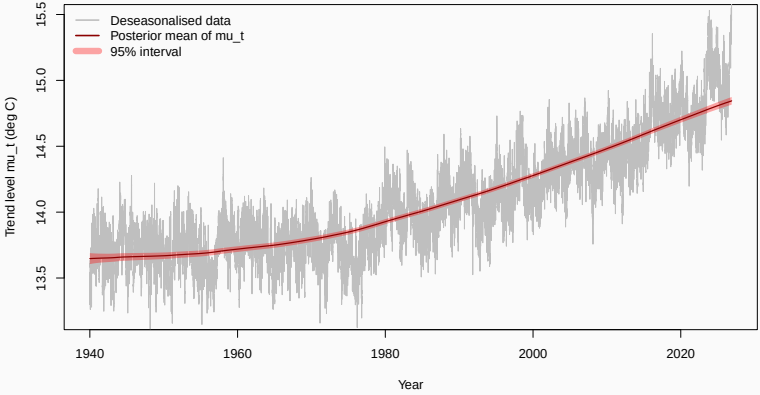
- **Month-of-year effects** $\gamma_1, \ldots, \gamma_{12}$, with $\sum_m \gamma_m = 0$: a static regression that captures *what is left* after the harmonics.
- **AR(2) short-term anomaly** (weather, ENSO-type persistence): $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \varepsilon_t$, $\varepsilon_t \sim N(0, rV)$.

In DLM form: $y_t = F_t'\theta_t + v_t$, $\theta_t = G\theta_{t-1} + w_t$, with a 21-dimensional state θ_t.

Prior, discounting and inference

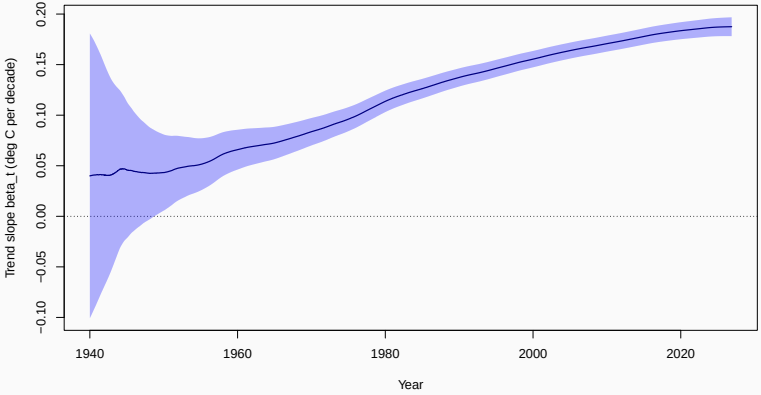
- **Unknown observational variance:** $V \sim IG(n_0/2,\, n_0 S_0/2)$ with $n_0 = 1$, $S_0 = 0.01$; vague $\theta_0 \sim N(m_0, C_0)$.
- **Component discounting** (West & Harrison, 1997, Ch. 6): W_t for the trend block $= \frac{1-\delta_T}{\delta_T} P_t^{\text{trend}}$, where $P_t = GC_{t-1}G'$; the same for the seasonal block with δ_S; month effects static ($W = 0$); AR innovation rV.
- **Exact conjugate updating**: Student-t one-step forecasts, filtering with $n_t = n_{t-1} + 1$ degrees of freedom, then retrospective smoothing (W&H Ch. 4) for $p(\theta_t \mid D_T)$. With $n_T \approx 31{,}700$, the posteriors are essentially normal.
- **Choices:**
  - $\delta_T = 0.9999$ (a long-run trend). Sensitivity to 0.999 and 0.99999 is shown later.
  - $\delta_S = 1$: when left free, the marginal likelihood pushed $\delta_S \to 1$, i.e. a *static* seasonal pattern.
  - (ϕ_1, ϕ_2, r) maximise the marginal likelihood $\sum_t \log p(y_t \mid D_{t-1})$ (type-II ML, i.e. empirical Bayes).
- Estimates: 1940 fit $\hat{\phi} = (1.49, -0.53)$, $\hat{r} = 52$; 1979 fit $\hat{\phi} = (1.51, -0.56)$, $\hat{r} = 42$. $\sqrt{\hat{V}} \approx 0.005\,^\circ$C: almost all short-term variation is in x_t.

Posterior of the trend level μ_t (1940–2026)



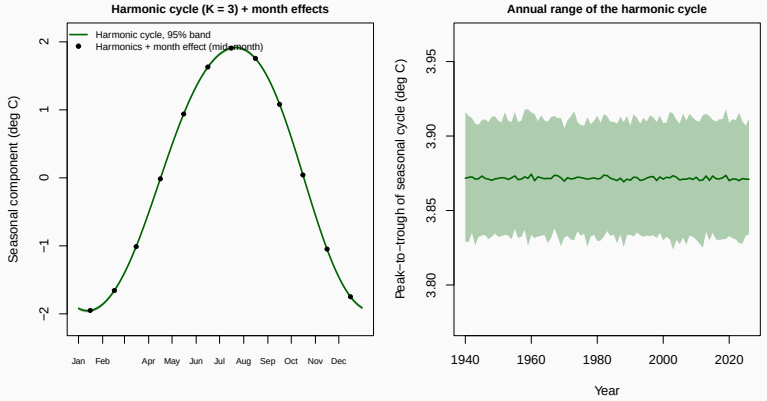
Smoothed posterior mean and 95% interval. Grey: data minus the seasonal components. μ_t goes from 13.65 °C (1941) to 14.84 °C (Nov 2026).

Posterior of the trend slope β_t (1940–2026)



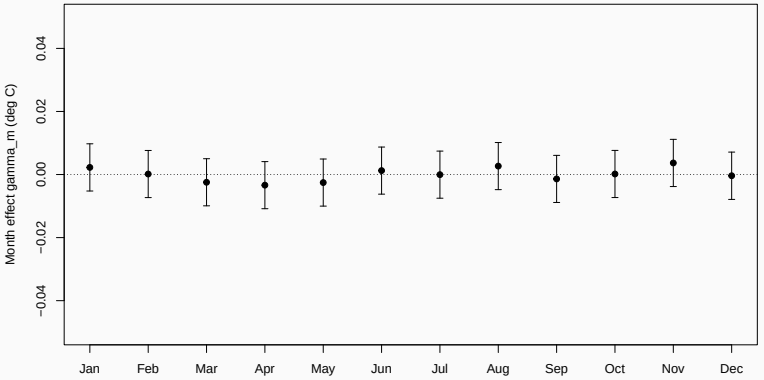
°C per decade: 0.07 (1960), 0.12 (1980), 0.16 (2000), 0.18 (2020), **0.19 [0.18, 0.20]** (2026). The warming rate is itself increasing.

Seasonality: harmonics and their stability (1940–2026)



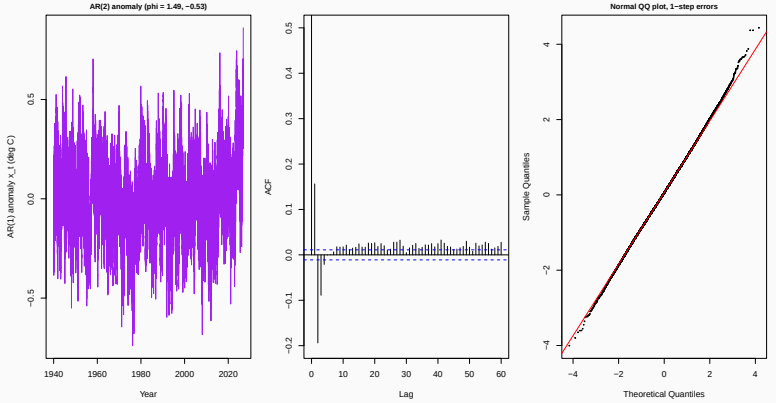
Left: posterior of the harmonic cycle; the dots add the month effects. Right: the peak-to-trough range, $\approx$**3.87 °C**, does not change over time (Monte Carlo from the smoothed posteriors).

What is left: the month-of-year effects (1940–2026)



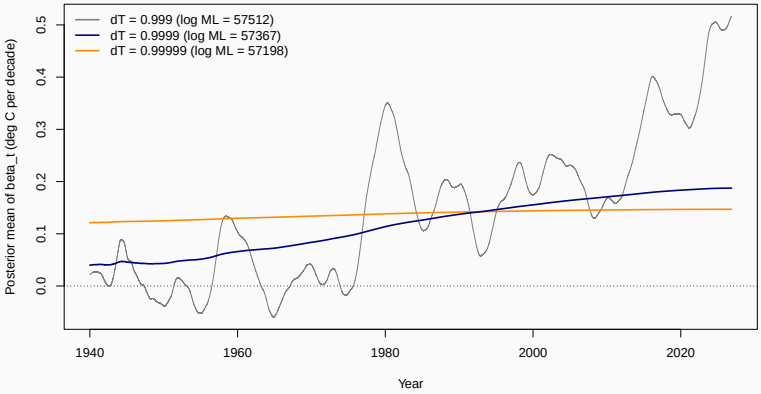
All $|\hat{\gamma}_m| < 0.005\,^{\circ}\mathrm{C}$, and every 95% interval contains 0. Three harmonics capture the global seasonal cycle, and the month dummies have nothing left to explain.

Short-term component and diagnostics (1940–2026)



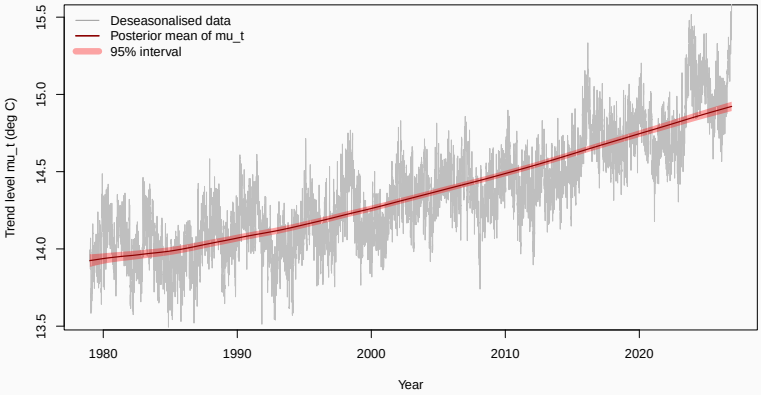
Some autocorrelation remains in the one-step errors (lag 1: 0.16; lag 2: -0.19). The AR(2) does not capture everything. The trend intervals are therefore probably *too narrow*, and the hyperparameters are plugged in, not integrated out.

Sensitivity to the trend discount factor δ_T (1940–2026)

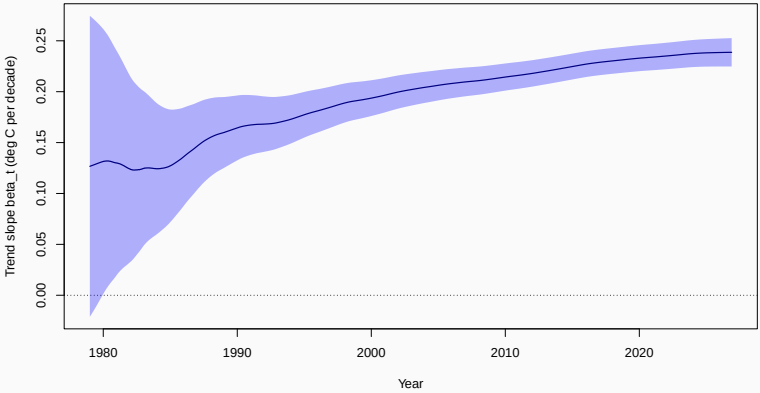


A smaller δ_T lets the slope chase multi-year swings (ENSO-like) and gets a higher marginal likelihood. A larger δ_T forces an almost constant slope. $\delta_T = 0.9999$ (a memory of decades) is a judgement call for a "long-run" trend.

The 1979–2026 fit: trend level

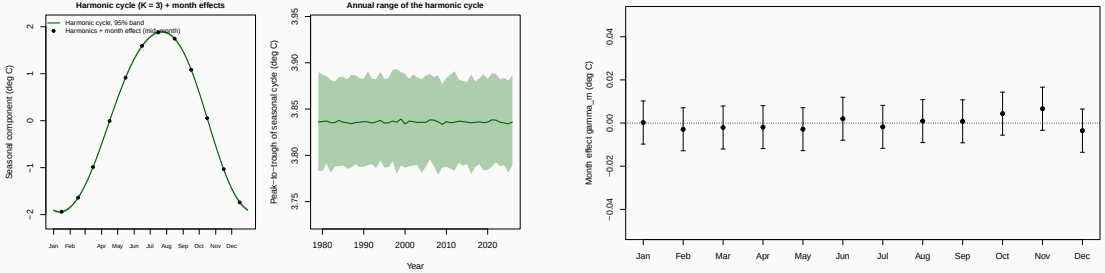


The 1979–2026 fit: trend slope



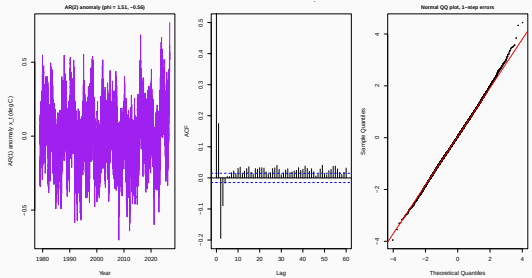
°C per decade: 0.13 [0.01, 0.25] (1980), 0.20 (2000), 0.23 (2020), **0.24 [0.22, 0.25]** (2026). This is higher than the 1940 fit for recent years, because the slow 1940–1970 period no longer anchors the trend.

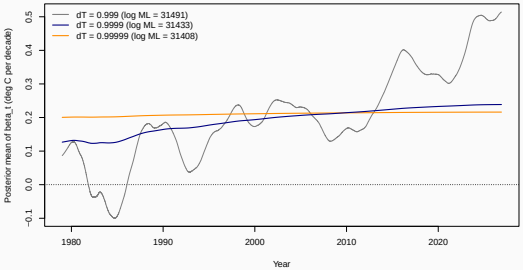
The 1979–2026 fit: seasonality and month effects



The range is 3.84 °C [3.78, 3.89], again constant over time. All month effects are within ±0.007 °C and their intervals contain 0.

The 1979–2026 fit: diagnostics and sensitivity





Residual autocorrelation at lags 1 and 2: 0.18 and -0.20. As before, a smaller δ_T gives a higher marginal likelihood but a slope that chases short-run swings.

SARIMA-type models for comparison

Classical counterparts

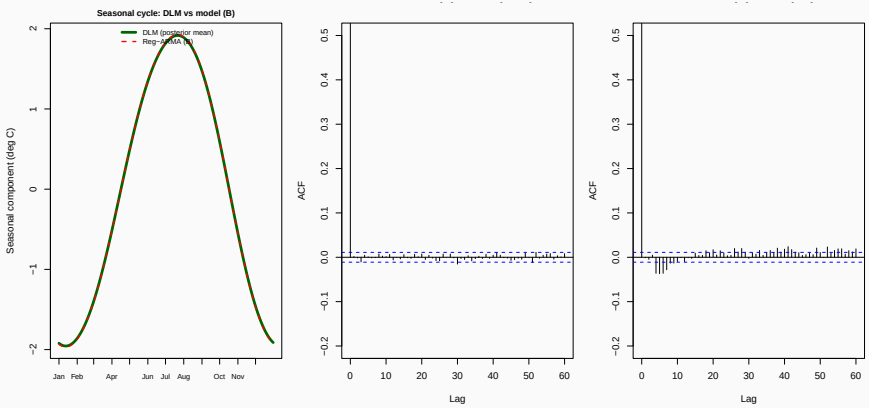
A seasonal ARIMA with period 365 is not practical with daily data, so the within-year seasonality enters as Fourier regressors, with the same $K = 3$ harmonics and 11 month effects as the DLM ("dynamic harmonic regression"):

$$y_t = \sum_{j=1}^{3} \left[c_j \cos(\omega_j t) + s_j \sin(\omega_j t) \right] + \gamma_{m(t)} + \delta\, t + \eta_t$$

- **(A)** $\eta_t \sim$ ARIMA(2,1,2), with drift δ: a stochastic level and constant slope, the closest analogue of the local linear trend.
- **(B)** $\eta_t \sim$ ARMA(3,1): a deterministic linear trend δt.
- **(C)** Monthly means: **SARIMA**$(1,0,1) \times (0,1,1)_{12}$ + drift, with a 24-month forecast.

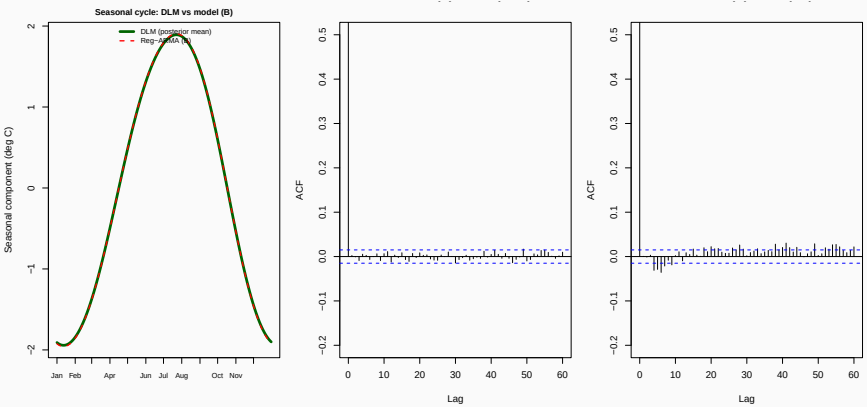
Orders were chosen by AIC among a few candidates, and all models were fitted by exact Gaussian ML (`stats::arima`).

Daily models (A) and (B), 1940–2026

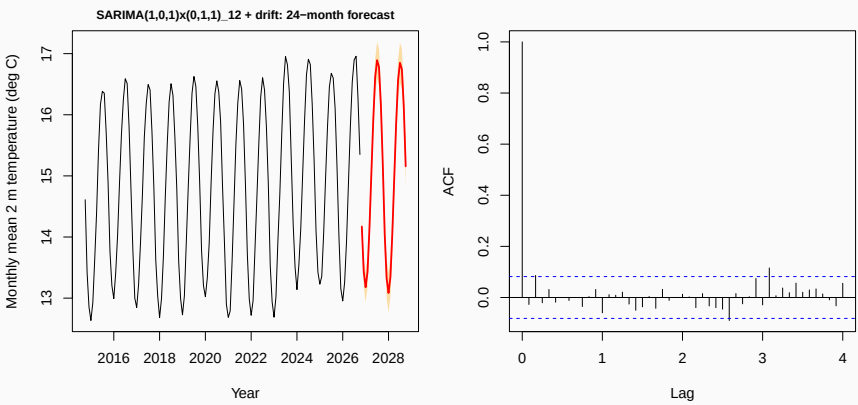


The seasonal cycle of (B) is indistinguishable from the DLM posterior mean. With MA terms, the residual autocorrelation is much smaller than for the DLM's AR(2).

Daily models (A) and (B), 1979–2026



Monthly SARIMA and a 24-month forecast (1979–2026 fit)



$\hat{\phi} = 0.91$, $\hat{\theta} = -0.37$, $\hat{\Theta} = -0.89$. Forecast to Oct 2028 with 95% bands; the residuals look close to white noise.

DLM vs SARIMA-type models

Start	Model	Slope (°C/decade)	1-step RMSE	1-step MAE	log pred. dens.
1940	DLM, β_t in 2026	0.19 [0.18, 0.20]	0.0387	0.0308	57,367
	DLM, average $(\mu_T - \mu_1)$/decades	0.14			
	(A) Reg-ARIMA(2,1,2) + drift	0.21 (se 0.21)	0.0366	0.0291	59,227
	(B) Reg-ARMA(3,1) + trend	0.14 (se 0.02)	0.0371	0.0295	58,841
	(C) Monthly SARIMA + drift	0.15 (se 0.03)	–	–	–
1979	DLM, β_t in 2026	0.24 [0.22, 0.25]	0.0384	0.0306	31,433
	DLM, average $(\mu_T - \mu_1)$/decades	0.21			
	(A) Reg-ARIMA(2,1,2) + drift	0.33 (se 0.30)	0.0361	0.0287	32,603
	(B) Reg-ARMA(3,1) + trend	0.22 (se 0.03)	0.0364	0.0290	32,424
	(C) Monthly SARIMA + drift	0.22 (se 0.04)	–	–	–

RMSE and MAE in °C, from one-step-ahead errors, excluding the first year. Log predictive density $= \sum_t \log p(y_t \mid y_{1:t-1})$ (DLM: Student-t marginal likelihood; ARIMA: Gaussian with plug-in MLE).

Reading the comparison

- **One-step fit:** the SARIMA-type models win slightly (RMSE 0.036 vs 0.038 °C). Their MA terms absorb short-term correlation that the DLM's AR(2) leaves behind. An ARMA component in the DLM would close the gap.
- **Trend:** the average slopes agree, about **0.14–0.15 °C/decade since 1940** and **0.21–0.22 °C/decade since 1979**. Only the DLM shows that the slope *itself increases*, with a posterior band at each date.
- The drift of (A) is very imprecise (se $\approx$ 0.2–0.3): with $d = 1$ it is essentially $(y_T - y_1)/T$.
- **Seasonality:** all models agree on a stable cycle with a range of about 3.9 °C. The month dummies add nothing beyond three harmonics.
- **2026:** the DLM puts most of this year's record warmth in the short-term component x_t, not in the trend.

When would we pass 2 °C? A forecasting exercise

Read this first: an exercise, not a climate projection

Disclaimer

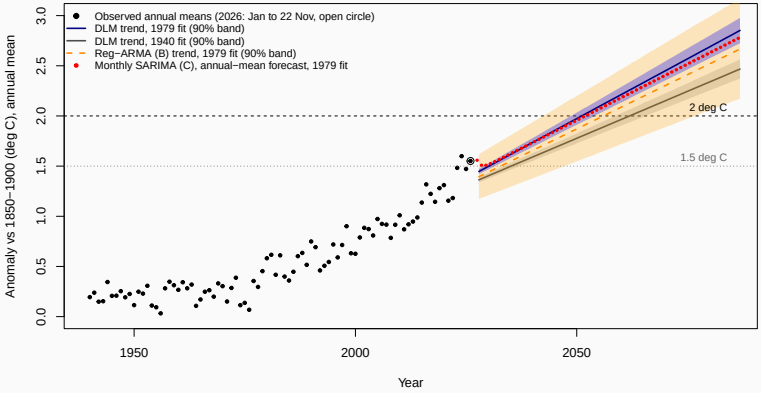
The numbers in this section are **not** a forecast of when the world will reach 2 °C. They are an **exercise in time series forecasting** with two of the most prominent *purely statistical* model classes: the Bayesian dynamic linear model and (S)ARIMA.

- These models know **nothing about physics**: no greenhouse gases, aerosols, emission scenarios, volcanoes, solar cycle or ocean heat uptake. They extrapolate the recent past.
- Every forecast here assumes **today's warming rate stays fixed**. The future rate depends on emissions decisions that no time series model can see.
- For real projections, see the physically based assessments: IPCC AR6 (2021), which uses climate models under explicit emission scenarios.
- **Do not quote these years at face value.** Use them to discuss what statistical extrapolation can and cannot say.

Setup

- Context: the global average surface temperature has been reported at about 1.47 °C above preindustrial, with 2026 among the warmest years on record (a strengthening El Niño plus ongoing warming).
- Target: annual mean **anomaly vs 1850–1900** $\geq 2\,^{\circ}\mathrm{C}$, with $A = \bar{y}^{\,\mathrm{t2m}} - \overline{\mathrm{clim}}_{1991\text{–}2020} + \overline{\mathrm{offset}}$ $(= \bar{y} - 14.38 + 0.88)$.
- **DLM:** 4,000 draws of $(\mu_T, \beta_T) \mid D_T$, propagated daily with the local-linear-trend evolution ($W_{T+k} = W_{T+1}$, discount $\delta_T = 0.9999$). The forecast slope stays near its 2026 value.
- **(B) reg-ARMA(3,1) + linear trend:** draws of (intercept, slope) from their asymptotic normal distribution.
- **(C) monthly SARIMA + drift:** 60-year forecast of the monthly means, averaged by year.
- "Trend crosses 2 °C" ≈ a 20-year mean crossing (the Paris definition). For "first single year above 2 °C" I add year-to-year noise with sd 0.10 °C: the scatter of annual means around a smooth trend since 1979, roughly the size of an El Niño boost.

Projected paths of the long-run trend



The trend in 2026 is 1.42 °C (DLM, 1979 fit) or 1.34 °C (DLM, 1940 fit). The reported 1.47 °C sits above the trend because of El Niño.

When does each model reach 2 °C?

Model	Trend now (°C)	Slope (°C/decade)	Trend crosses 2 °C	First single year > 2 °C
DLM, 1979 fit	1.42	0.24	∼2052 (2049–2055)	∼2048 (2043–2053)
(B) reg-ARMA, 1979 fit	1.37	0.22	∼2056 (2043–2075)	∼2052 (2039–2071)
(C) monthly SARIMA, 1979 fit	–	0.22	annual-mean forecast:	∼2052 (2038–2068)
DLM, 1940 fit	1.34	0.19	∼2063 (2059–2066)	∼2058 (2051–2064)
(B) reg-ARMA, 1940 fit	1.20	0.15	2070s or later	–

90% intervals in brackets. For (C), the bracket runs from the first year with Pr(annual > 2) $\geq 5\%$ to the first with $\geq 95\%$.

- So the answer to "2035, 2040 or 2045?" is **none of them**. The constant-slope extrapolations put the trend crossing around **2050** (1979 fits), or the 2060s (1940 fits).
- A single warm year could cross a few years earlier, mostly in the 2040s.

Why not to take these numbers at face value

1. **Constant slope.** The DLM itself shows the slope doubled since 1980 ($0.12 \rightarrow 0.24\,^{\circ}\mathrm{C}$/decade). If the acceleration continues (e.g. $+0.025\,^{\circ}\mathrm{C}$/decade per decade), the crossing moves to $\approx$2048. If emissions fall, it moves later.
2. **Choice of start date.** 1940 vs 1979 shifts the answer by about a decade: the model sees a different past and extrapolates it.
3. **Overconfident DLM intervals.** The hyperparameters are plugged in, and some residual autocorrelation remains. The SARIMA intervals are wider and more honest about this.
4. **60-year horizons.** No statistical model fitted to 47–87 years of data should be trusted that far ahead.
5. **Physical benchmark.** As far as I know, IPCC AR6 puts the $2\,^{\circ}\mathrm{C}$ crossing (20-year mean) roughly in 2041–2060 under an intermediate-emissions scenario. That agrees with the 1979-based fits, possibly by coincidence.

Teaching point: the models agree closely on the *past* trend and disagree substantially about the *future*. That gap is mostly model structure and start date, not sampling noise.

Wrap-up

Take-aways

1. Copernicus Climate Pulse (`https://pulse.climate.copernicus.eu`) makes ERA5 daily global temperature since 1940 freely available: one CSV, updated daily.
2. ERA5 is a *reanalysis*: a gap-free, area-weighted model estimate built from observations. It is robust from 1979; earlier decades rely more on the model.
3. The warming trend is about **0.19 °C/decade now** (1940 fit) or **0.24 °C/decade** (1979 fit), up from ≈0.07 in 1960.
4. The global seasonal cycle (≈3.9 °C, peaking in July because of Northern Hemisphere land) is stable. Three harmonics are enough, and the month effects are zero.
5. The DLM and SARIMA-type models agree on the average trend. The DLM adds an interpretable, time-varying slope with posterior uncertainty. SARIMA has slightly better one-step accuracy.
6. As a forecasting exercise only, constant-slope extrapolation puts the trend at 2 °C around 2050 (1979 fits) or the 2060s (1940 fits). This is not a climate projection.

Possible extensions: full Bayesian treatment of (ϕ, r, δ) by MCMC; ARMA or ENSO covariates in the DLM; stochastic volatility in v_t; regional series from the CDS.

References and code

- Copernicus Climate Change Service, *Climate Pulse*. https://pulse.climate.copernicus.eu
- Hersbach, H. et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146, 1999–2049.
- West, M. & Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd ed. Springer.
- Prado, R., Ferreira, M. A. R. & West, M. (2021). *Time Series: Modeling, Computation, and Inference*, 2nd ed. CRC Press.
- IPCC (2021). *Climate Change 2021: The Physical Science Basis*. Contribution of Working Group I to the Sixth Assessment Report. Cambridge University Press.
- Box, G. E. P., Jenkins, G. M., Reinsel, G. C. & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*, 5th ed. Wiley.

R code (base R only), in this order: `01_descriptive_plots.R`, `dlm_functions.R` + `02_dlm.R` (W&H filter and smoother written from scratch), `03_sarima.R`, `04_crossing_2C.R` + `05_crossing_plot.R` (2 °C exercise).