Why Forecasts Go Wrong

By the rock. August 25, 2026. Grounded in National Weather Service definitions.

The short version

Forecasts go wrong because the atmosphere amplifies small errors. No set of observations is perfect or complete, and the equations of weather compound every imperfection, doubling small differences until two nearly identical starting points produce two different weekends. That is chaos, demonstrated by Edward Lorenz in 1963, and it is why forecasters now run ensembles: the same model dozens of times with slightly wobbled starts, reading confidence from the spread. You can watch the decay in this site's own graded record: as of the receipts page's August 13 update, our wet-or-dry calls verified 92 percent of the time one hour out and 77 percent at 48 hours. Same rock, same math, longer lead.

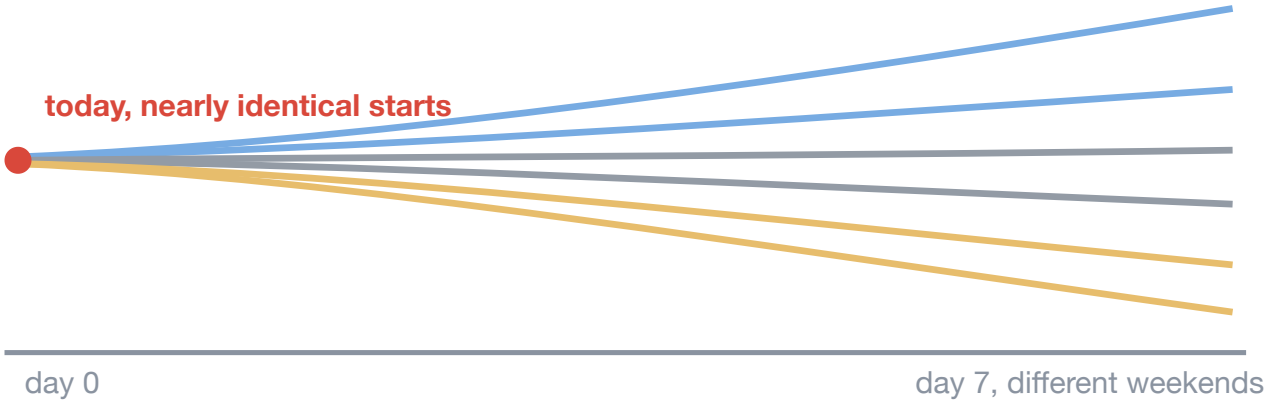
The rerun that would not repeat

In 1961, an MIT meteorologist named Edward Lorenz was running a small weather model, twelve equations, on a Royal McBee LGP-30 computer. Wanting to re-examine one stretch, he restarted the run from the middle, typing in the numbers from a printout. The printout rounded to three decimal places: he entered 0.506 where the machine had been carrying 0.506127. A difference of about one part in ten thousand, far smaller than any real weather instrument can even measure.

The rerun tracked the original for a while, then drifted, then bore no resemblance to it at all. Lorenz had discovered, by accident and then confirmed with great care, that the equations of the atmosphere amplify tiny differences instead of forgetting them. He published the result in 1963 as Deterministic Nonperiodic Flow, in the Journal of the Atmospheric Sciences, and it became one of the founding papers of chaos theory. The practical sentence inside the mathematics: even a perfect model of the atmosphere, fed almost perfect observations, produces a forecast that degrades with time, because almost is a load-bearing word.

Error grows on compound interest

Every forecast starts from a snapshot of the atmosphere built from surface stations, balloons, aircraft, and satellites, and the snapshot has holes: instruments have tolerances, oceans are sparsely sampled, and the air between observations is estimated. Those small initial errors do not sit still. The atmosphere's own physics multiplies them, the errors infect neighboring regions, small-scale mistakes grow into large-scale ones, and past a week or two the compounding swamps the signal for day-to-day details. This is not a hardware problem or a laziness problem. Better models and better observations push the horizon outward, and famously have: a widely cited 2015 review in Nature called it the quiet revolution, forecast skill gaining roughly one day of usable lead time per decade. But the horizon moves; it does not go away. Chaos sets the terms.



An ensemble, schematically: one model, many almost identical starting points. Where the lines hug each other, trust the forecast. Where they fan out, the honest answer is the fan.

Ensembles: wobble it and run it again

Since the starting snapshot cannot be perfect, modern forecasting stopped pretending it was. Forecast centers run the same model dozens of times, each run beginning from a slightly different, equally plausible version of right now. The bundle is called an ensemble, and it turns chaos from an enemy into an instrument. When the runs agree five days out, the atmosphere is in a forgiving mood and confidence is high. When they scatter, the spread itself is the forecast: it says the answer is genuinely not knowable yet, and says it honestly. Every chance of rain you have ever read is downstream of this machinery; that page covers what the percent does and does not promise.

Our own miss record, in public

Most weather sites tell you they are accurate. This one keeps a receipts page, where every forecast the rock publishes is written down, waited on, and graded against what actually happened, and the numbers print whether they flatter the rock or not. It is the primary source for this lesson, and as of its August 13, 2026 update, with 2,827 graded calls on the ledger, it shows error growth in the wild. Wet or dry, our calls verified 92 percent one hour ahead, 91 percent at six hours, 88 percent at 24, and 77 percent at 48. Temperature misses, meanwhile, barely budged: off by 2.4 degrees at one hour and 2.6 at 48. Lead time punishes the yes-or-no questions first.

Two smaller lessons hide in the same table. First: the 12-hour row reads 96 percent, better than the one-hour row, which should smell wrong, and mildly is. That row holds 175 calls; a sample that size wobbles, which is why the receipts page flags thin samples instead of celebrating them. Second: on the day-ahead rain question, the three professional models we grade did not tie. Over roughly 600 days each, one verified 71 percent, the others around 60. Different models, same atmosphere, different errors: that disagreement is exactly the spread an ensemble is built to measure.

Where forecasts actually break

Forecasts rarely fail everywhere at once; they fail by category. Temperature usually misses small, a degree or three, and mostly from local effects. Timing misses medium: the front arrives, just four hours late, and your dry morning was real while your dry afternoon was not. Precipitation placement misses

biggest, especially on scattered-storm days when the atmosphere itself has not decided which county wins. Knowing WHICH kind of miss to expect is most of the skill of using a forecast.

Grade one week yourself

Tonight, write down tomorrow's forecast for your spot: high temperature, sky, whether it rains, and when. Tomorrow night, grade it against what happened, in three columns: temperature, timing, precipitation. Repeat for seven days. One week is a thin sample, as the 96 percent row just taught you, but the pattern still shows: the temperature column will mostly hold, and the misses will pool in timing and precipitation. By Sunday you will have built a miniature receipts page of your own, and you will read every forecast afterward the way forecasters do: as odds, with a known weak spot.

Check your understanding

1. What did Lorenz's 1961 rerun demonstrate?

- □ A. His computer was broken
- □ B. Tiny differences in starting conditions grow into completely different weather
- □ C. Weather models need more than twelve equations

2. Why do forecast centers run the same model dozens of times with slightly different starts?

- □ A. To keep the computers busy between storms
- □ B. Because the starting snapshot cannot be perfect, and the spread of results measures confidence
- □ C. To average away the need for observations

3. On this site's own graded record, what happened to wet-or-dry accuracy between 1 hour and 48 hours of lead?

- □ A. It held steady near 92 percent
- □ B. It fell from 92 percent to 77 percent
- □ C. It rose, because longer forecasts are more careful

4. The 12-hour row on the receipts page reads 96 percent, better than the 1-hour row. The right reaction is:

- □ A. Forecasts are most accurate at exactly 12 hours
- □ B. Suspicion: 175 calls is a small sample, and small samples wobble
- □ C. The 1-hour forecasts must be broken

5. Which part of a day-ahead forecast typically misses by the most?

- □ A. Temperature
- □ B. Precipitation placement and timing
- □ C. The date

Answer key

1. What did Lorenz's 1961 rerun demonstrate?

B. Tiny differences in starting conditions grow into completely different weather — He re-entered rounded numbers, one part in ten thousand off, and the rerun diverged from the original entirely. The atmosphere's equations amplify small errors rather than forgetting them; he published the result in 1963.

2. Why do forecast centers run the same model dozens of times with slightly different starts?

B. Because the starting snapshot cannot be perfect, and the spread of results measures confidence — That is an ensemble. Agreement among the runs means the atmosphere is in a predictable regime; scatter means the answer is genuinely not knowable yet, and the spread says so honestly.

3. On this site's own graded record, what happened to wet-or-dry accuracy between 1 hour and 48 hours of lead?

B. It fell from 92 percent to 77 percent — The receipts page's August 2026 update grades 92 percent at one hour and 77 percent at 48. That decay is error growth measured on real published forecasts, ours.

4. The 12-hour row on the receipts page reads 96 percent, better than the 1-hour row. The right reaction is:

B. Suspicion: 175 calls is a small sample, and small samples wobble — Skill cannot genuinely improve with more lead. A thin sample bounces around its true value, which is why the receipts page flags small samples instead of bragging with them.

5. Which part of a day-ahead forecast typically misses by the most?

B. Precipitation placement and timing — On our own ledger, temperature misses stayed near two and a half degrees at every lead while wet-or-dry accuracy fell steadily. The yes-or-no and the when are what lead time eats first.

Questions people actually ask

If chaos limits forecasts, why have they gotten so much better?

Better observations, better models, and better use of ensembles keep shrinking the initial errors and squeezing more out of the physics. A 2015 review in Nature put the pace at roughly one extra day of usable lead time per decade. Chaos does not forbid improvement; it sets a horizon that improvement approaches expensively.

Is a forecast that said 20 percent chance of rain wrong when it rains?

No. A 20 percent day rains one time in five, and the forecast priced that in. What the percent means, and how to hold it accountable fairly, is the chance of rain lesson; how often OUR percents kept their promises is a table on the receipts page.

Why do different weather apps disagree about the same day?

They are usually reading different models, or blending them differently, and the models genuinely disagree; on our own day-ahead grading the three we track ran from 71 percent down to 59. When apps disagree, you are looking at ensemble spread with branding.

What is the hardest thing to forecast?

Among everyday weather: exactly where and when scattered storms fire. The setup is predictable; the specific county is often not, because storm-scale details grow from below what observations resolve. That is chaos operating at its fastest scale.

Does this site really grade its own forecasts?

Every one it publishes, against independent observations of what actually happened, with the misses printed alongside the hits. The receipts page updates as the ledgers grow, so its numbers will drift from the snapshot quoted in this lesson; the drift is the point.
