AN AGENT LOOP FOR AUTONOMOUS, REASONING-DRIVEN TUNING OF NUMERICAL WEATHER PREDICTION MODELS

D. Craggs, P. D'Abreton, A. Twidle
Ideagen Envirosuite

Abstract

Configuring the Weather Research and Forecasting (WRF) model for accurate near-surface meteorology is time- and expertise-intensive, and no single configuration performs well across different climates. We test whether an autonomous agent can do this work: selecting and refining WRF's physics schemes and model structure for a site, from observations alone, the way an expert would. We present a multi-agent harness built on shared on-disk state. A reasoning agent proposes configurations using knowledge held in the model weights, targeted web search and a curated set of rules and lessons; an execution agent runs the resulting WRF simulations in parallel batches, which are scored against surface observations and fed back into the next round of reasoning. Given only a site and its observations, and with no human input during a campaign, the harness improves the configuration toward the Emery et al. (2001) surface-meteorology benchmarks across four climatically distinct Australian sites. Two sites reach all eight Emery benchmarks for 2 m temperature, 10 m wind speed and wind direction; the other two reach seven of eight, each missing on a single wind metric. The agent arrives at these configurations by diagnostic, hypothesis-led reasoning over discrete scheme and structural choices rather than a pre-set grid or sweep. These results show that autonomous, reasoning-driven configuration can take WRF to, or within one benchmark of, an established surface-meteorology standard.

Keywords: WRF, autonomous agents, model configuration, surface meteorology

1. Introduction

Accurate near-surface meteorology underpins much of atmospheric science, from site-specific forecasting to air-quality and dispersion modelling. The WRF model (Skamarock et al. 2019) is a powerful tool for producing it, downscaling large-scale synoptic conditions to the regional and local effects that shape surface weather. Its strength is versatility, a wide range of physical and structural options, but that same versatility sets a high bar: configuring the model well for a site takes considerable knowledge of the extensive physics parameterisation options available and hands-on experience.

The space of valid configurations is enormous: the physics schemes and model settings combine in far more ways than any practitioner could test directly. Experience, the literature and established physics suites such as the CONUS (contiguous United States) suite narrow it, but it stays broad. Reaching a strong configuration is therefore an iterative, expert-driven process (Duan et al. 2017): a practitioner picks a plausible starting point, runs it, reasons over where it falls short, adjusts, and runs again, working toward the best result they can reach.

Automated tuning offers another route. Surrogate-assisted optimisation has delivered measurable forecast gains in WRF itself (Duan et al. 2017), and Bayesian calibration and history matching are established ways to tune climate-model physics (Couvreux et al. 2021). Such searches are effective at exposing how settings relate and how sensitive a model is to them, but the parameters and their ranges must be chosen up front, and the exercise tends to reveal more about the tested parameters themselves than about the best configuration for one site.

Outside weather prediction, autonomous agents have begun to play a substantial role in scientific work, from systems that run an entire research process (Lu et al. 2026) to narrower ones that drive an existing solver toward a target, such as a large language model (LLM) agent steering OpenFOAM for computational fluid dynamics (Pandey et al. 2025). We ask whether such an agent can do what an experienced weather-forecasting practitioner does, tuning WRF by reasoning over results and experience rather than by sweeping a fixed space, and without supervision.

We answer with an autonomous loop that runs the process end to end. An LLM, given research tools and a curated base of rules and lessons, proposes WRF configurations, runs them and scores them against observations, and iterates. We test it across four climatically distinct Australian sites, each over a month-long campaign. The agent improves through each campaign and shows diagnostic reasoning that targets the right deficiency at each site, reaching all

eight Emery benchmarks at two sites and seven of eight at the other two.

2. An agent loop for WRF configuration

Progress in LLMs over the past few years has come not only from raw model capability but from the harnesses around them, most visibly in software engineering and increasingly in scientific research. The key primitive is the agent: an LLM given tools and the latitude to act with some autonomy, which can propose an experiment, run it, read the result and choose the next move. That turns assistance into execution. A productive way to harness the behaviour is the agent loop, popularised in early 2026 by Karpathy's autoresearch, which set a coding agent loose on hundreds of unattended experiments against a plain-text goal and a fixed per-experiment budget (Karpathy 2026). The loop has since become a common primitive, shipping as a built-in goal-seeking mode in leading agent harnesses including Anthropic's Claude Code and OpenAI's Codex.

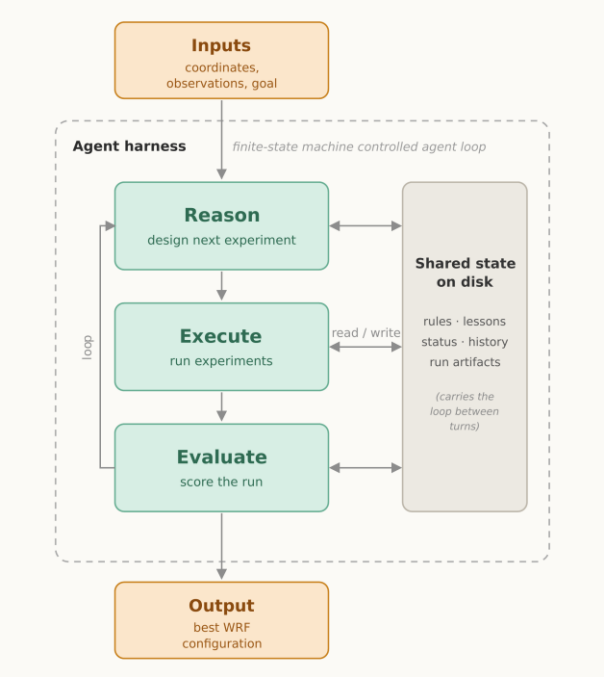


Figure 1. The agent harness. Inputs (site coordinates, observations and the goal) enter a finite-state machine that cycles three roles, each a fresh language-model session: reason designs the next experiment and calls web search for literature, execute runs the scripts and WRF, and evaluate scores the run against Emery. All roles read and write shared state on disk (rules, lessons, status, history and run artifacts), which carries the loop from one turn to the next. The output is the best configuration found.

This paper applies that loop to WRF configuration. The harness (Fig. 1) has the parts such a system needs: a defined input, the site and a goal; separate roles for designing, executing and evaluating experiments; external tools and data the session can call; a curated base of rules and lessons that bounds the search; and a shared state on disk that persists outside any single session.

Concretely the harness is a finite-state machine cycling three roles, reason, execute and evaluate. Each role runs as its own agent session, and that session ends at the close of every pass: history and logs are written to shared state on disk, and the next role resumes in a new session with a clean context window. This boundary is by design: it keeps context and token use in check and binds the agent's attention to the latest result and the goal as written on disk rather than to a long, drifting transcript. It also makes the machine robust to being paused and resumed, by intent or after a failure, since nothing lives only in a session and a broken run simply restarts.

The goal is fixed at the outset: optimise the WRF configuration toward a set of surface-meteorology benchmarks, scored against the site's observations. From there the agent proposes configurations, runs and scores them, and needs no human guidance during a campaign.

3. Experimental setup

The four sites span distinct Australian climates: the Gulf Country, QLD (semi-arid inland), the Pilbara, WA (hot desert), the Peel region, WA (coastal Mediterranean), and the Bowen Basin, QLD (wet-season subtropical inland). Periods and observation details are in Table 1.

Table 1. The four campaigns. Scored periods are in local standard time (AEST for Queensland, AWST for Western Australia) and begin about a day before the target month for spin-up; paired hours are valid hourly station-to-model pairs. Shared setup: WRF-ARW 4.6.0, three one-way nests at 18 / 6 / 2 km, 45 eta levels, 34 GFS metgrid levels, GFS forcing, nearest-grid pairing.

Region	Climate regime	Scored period	Paired hours
Gulf Country, QLD	Semi-arid inland	30 Jun – 1 Aug 2025	768
Pilbara, WA	Hot desert	30 Jun – 1 Aug 2022	768
Peel region, WA	Coastal Mediterranean	31 Dec 2025 – 1 Feb 2026	768
Bowen Basin, QLD	Wet-season subtropical inland	31 Jan – 1 Mar 2026	696

Each campaign runs to a fixed 168-hour wall-clock budget on a cloud instance, up to four WRF experiments at a time. Forcing is the GFS analysis at forecast hour 000 from each cycle (NCEP 2003), not a reanalysis and not a free forecast; the runs sit between a reanalysis-forced hindcast and an operational forecast, and because every experiment in a campaign uses the same forcing it does not affect the comparison of configurations. All campaigns share the model setup in Table 1; the physics schemes are what the agent selects per site

(Section 4), and the reasoning model, reasoning effort and compute are in Table 2.

Table 2. Agent and compute.

Component	Choice
Model provider	Anthropic
Model	Claude Opus
Reasoning effort	xhigh
LLM harness	Claude Code
Compute	AWS EC2 (c8i.16xlarge)
WRF build	Compiled from source in a Docker image

We adopt the Emery et al. (2001) surface benchmarks as the scoring standard, a widely used reference for meteorological model evaluation; runs are scored on eight scalars covering 2 m temperature, 10 m wind speed and 10 m wind direction. Each site is evaluated against a single surface station, with observations from the Bureau of Meteorology or, with permission, from Envirosuite users' stations, paired to the nearest grid centre (for temperature) and grid sides (for wind components) on the inner domain. Above-surface, multi-station evaluation and interpolated pairing are out of scope.

4. Results

Table 3 summarises the four campaigns. All four ended at or within one benchmark of the full Emery set: two configurations meet all eight surface scalars (the Gulf Country and the Bowen Basin, both QLD) and two meet seven of eight (the Pilbara and the Peel region, both WA). The pace varied, from the Gulf Country reaching a strong configuration in a handful of batches to the Pilbara and the Peel region taking many more.

Table 3. Outcome per campaign. Batches and wall-clock to the best configuration, against a 168-hour budget; campaigns continued exploring within budget after a first pass.

Campaign	Batches	Time (h)	Best configuration	Emery
Gulf Country, QLD	5	37	QNSE PBL, Noah land surface, RRTMG radiation	8 / 8
Pilbara, WA	17	129	MYJ PBL, above-PBL nudging on the inner domain	7 / 8
Peel region, WA	11	164	MYNN2.5 PBL, Noah-MP, low-level eta stretch	7 / 8
Bowen Basin, QLD	9	103	ACM2 PBL, five-layer slab land surface	8 / 8

Table 4 gives the full Emery statistics for each best configuration. The two seven-of-eight results each miss on one wind metric: the Pilbara on wind-direction gross error in light winds, the Peel region on a nocturnal 10 m wind-speed bias. Figure 2 places each best configuration on Taylor diagrams against the observed hourly series: temperature skill is high at every site, while wind is the harder field.

Table 4. Emery surface statistics for the best configuration. Columns are the four campaigns; rows are the Emery surface metrics. Thresholds: T2m gross error ≤ 2.0 °C, bias ≤ ±0.5 °C, IOA ≥ 0.8; WS10 RMSE ≤ 2.0 m/s, bias ≤ ±0.5 m/s, IOA ≥ 0.6; WD10 gross error ≤ 30°, bias ≤ ±10°. Bold exceeds threshold.

Metric	Gulf Country, QLD	Pilbara, WA	Peel region, WA	Bowen Basin, QLD
T2m bias	-0.05	-0.04	-0.01	0.06
T2m GE	1.51	1.58	1.51	1.24
T2m IOA	0.97	0.98	0.98	0.95
WS10 bias	-0.09	0.11	**0.93**	0.26
WS10 RMSE	0.91	0.97	2.00	1.32
WS10 IOA	0.89	0.93	0.74	0.83
WD10 bias	1.1	-6.4	5.8	3.1
WD10 GE	19.5	**33.6**	28.8	22.7
/8	8	7	7	8

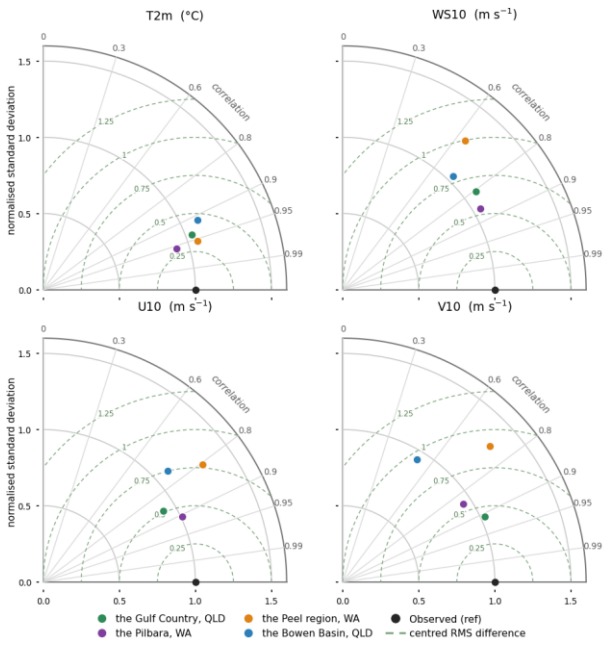


Figure 2. Taylor diagrams from the recovered hourly series. Correlation (angle) and normalised standard deviation (radius) against the observed reference (the dark point). Temperature skill is high everywhere; wind is harder, with direction shown through its U and V components, since a direction Taylor is not statistically well-defined, and wind speed shows mild over-variability at most sites.

5. How the configurations were reached

Across the campaigns some metrics were quick to satisfy while others took sustained work. To move toward passing all eight Emery benchmarks, the agent worked on whichever metric was currently furthest from its target, the binding constraint. This mirrors how a campaign is scored: a configuration passes only when its worst metric falls below its own

Emery threshold, so progress is set by that single worst metric rather than by a summed distance to perfection that a few easy metrics can flatter. Figure 3 traces the binding constraint against wall-clock for each campaign. All four descend toward the pass line. Two cross it: the Gulf Country early, by its fifth batch, after which it keeps exploring within budget, and the Bowen Basin at its final batch. The Pilbara and the Peel region descend toward the line and flatten just above it, each ending one metric short.

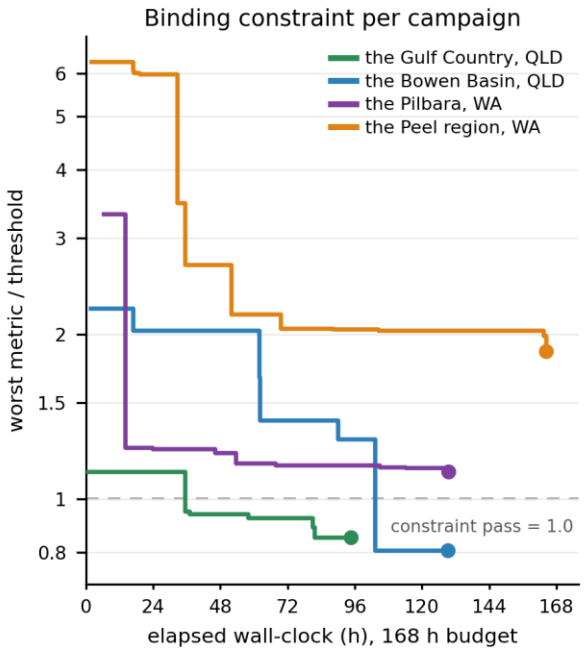


Figure 3. The binding constraint over wall-clock. For each campaign, the worst surface metric relative to its Emery threshold, on a log axis with the pass line at 1.0, against elapsed wall-clock hours. The Gulf Country and the Bowen Basin cross below the line (eight of eight); the Pilbara and the Peel region descend and flatten just above it (seven of eight).

Behind that single constraint lies a large configuration space. The agent works from an inventory of sixteen discrete physical and structural levers, among them the boundary-layer and surface-layer schemes, the land-surface model, radiation, microphysics, cumulus, vertical structure and nudging, whose valid combinations make up the enormous space noted in the introduction. Against that scale the four campaigns scored 133 experiments in total, a few dozen per site.

Figure 4 shows the extent and variability of that search: the number of distinct settings the agent tried for each lever across the four campaigns, grouped by physics family. Exploration concentrated on the choices that most affect near-surface meteorology: the boundary-layer scheme drew the most variants, followed by the vertical level structure, the land-surface model, the surface-layer scheme and the number of soil layers. Microphysics, dynamics and the radiation cadence were varied far less, consistent with their weaker influence on 2 m temperature and 10 m wind over these short, surface-focused campaigns. The shape of the figure is itself evidence of diagnostic search rather than a uniform sweep: a sweep would touch every lever roughly equally, whereas the agent invested where the surface metrics were most sensitive.

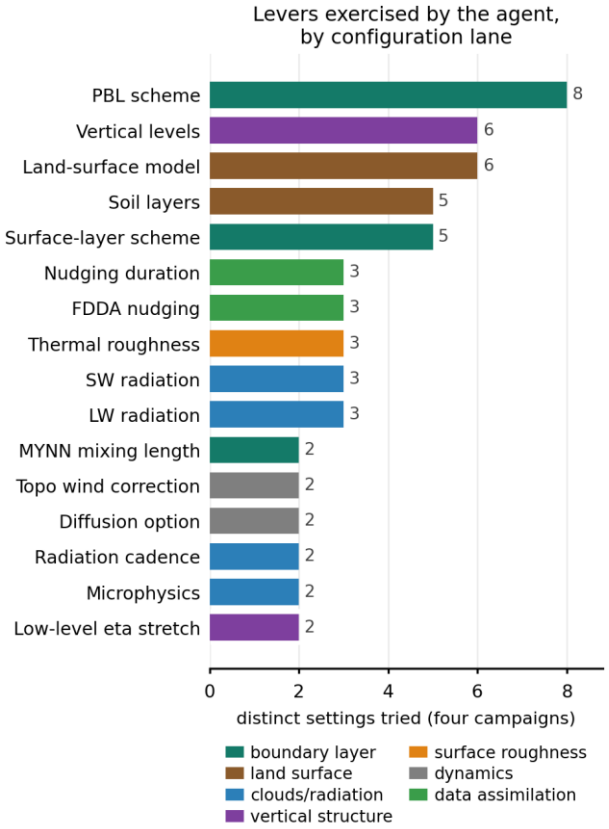


Figure 4. Levers exercised by the agent. The number of distinct settings the agent tried for each of the sixteen configuration levers across the four campaigns, coloured by physics family. Effort concentrated on the boundary-layer, vertical-structure and land-surface choices that most affect near-surface meteorology, with far fewer settings spent on microphysics, dynamics and radiation.

The decision record shows this as diagnostic work rather than a uniform procedure. In the Peel region the binding constraint was the 2 m temperature error, which sat above its threshold for most of the campaign. In its eleventh batch the agent ran a two-by-two factorial that crossed a low-level vertical-grid stretch (the automatic spacing against roughly ten levels in the lowest 450 m) with a surface-resistance option, then read the result: the grid stretch was the dominant lever, lowering the temperature error mainly in the early-night inversion hours, consistent with terrain-driven nocturnal drainage where the site lies on a coastal plain below an escarpment to the east, and additive with a smaller gain from the resistance change. That batch produced the campaign's first temperature pass and took the Peel region to seven of eight, leaving a single residual in the 10 m wind-speed bias. Because the harness keeps the full record, this change and the others are

traceable to the residual each was meant to fix. We do not claim the search is optimal, only that it is recorded and open to inspection.

6. Discussion and conclusion

The harness did what it was built to do: tune WRF autonomously, reasoning over scheme and structural choices toward the Emery surface benchmarks. Skill improved over the course of each campaign, and the final configurations met all eight benchmarks at two sites and fell one metric short at the other two. The agent did this from a cold start, with no seed configuration and no human input during the campaigns. We chose that unaided setting deliberately, to measure how far the agent gets on its own rather than because it is the way the system is meant to be used; the result is best read as a floor on capability, not a ceiling.

Realistic use would be collaborative rather than solo. A practitioner would seed the agent with configurations known to work in similar climates and steer it as it runs, approving directions or redirecting the search, which would likely reach stronger configurations or reach them sooner. The lever set could also widen. Both residual misses are wind metrics whose plausible fixes sit outside the space the agent was allowed to search: the Peel region's nocturnal wind-speed bias points at surface roughness and drainage flows that updated land-use data or a finer inner nest could resolve, and the Pilbara's light-wind direction error is as much a forcing question as a physics one, where ERA5 reanalysis boundaries may help. Extending the agent's reach to these levers is the same loop over a larger space.

These results carry clear limits: four short (1-month) campaigns, each covering a single season rather than a paired summer and winter, a single surface station at each site, no controlled grid or random-search baseline over the same lever space, and the levers left out of scope above. Within those limits, autonomous, reasoning-driven configuration of WRF for near-surface meteorology is feasible, and it can be done with a complete, inspectable record of how each configuration was reached. The contribution is not an argument for removing the practitioner but a demonstration that the agent can carry the search itself, and a basis for using it as a collaborator that a practitioner steers.

References

Couvreux F. et al. 2021, 'Process-based climate model development harnessing machine learning: I. A calibration tool for parameterization improvement', Journal of Advances in Modeling Earth Systems.

Duan Q. et al. 2017, 'Automatic model calibration: a new way to improve numerical weather forecasting', Bulletin of the American Meteorological Society.

Emery C., Tai E. & Yarwood G. 2001, Enhanced meteorological modeling and performance evaluation for two Texas ozone episodes, report, ENVIRON/TNRCC.

Karpathy A. 2026, autoresearch, GitHub, https://github.com/karpathy/autoresearch.

Lu C. et al. 2026, 'Towards end-to-end automation of AI research', Nature, vol. 651, pp. 914–919.

NCEP 2003, The GFS atmospheric model, NCEP Office Note 442.

Pandey S. et al. 2025, 'OpenFOAMGPT: a retrieval-augmented large language model (LLM) agent for OpenFOAM-based computational fluid dynamics', Physics of Fluids, vol. 37, 035120.

Skamarock W.C. et al. 2019, A description of the Advanced Research WRF model version 4, NCAR Technical Note.