Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When do inflow forecasts add value? Closed-loop evidence on foundation models, hedging and scarcity in the Brazilian hydrothermal system

leonardo Forero Mendoza

mendonza@ica.ele.puc-rio.br

Universidade do Estado do Rio de Janeiro

Harold D. de Mello Jr.

Universidade do Estado do Rio de Janeiro

Antonio Guilherme Garcia Lima

Universidade do Estado do Rio de Janeiro

Marco Aurelio Pacheco

Pontifícia Universidade Católica do Rio de Janeiro

Research Article

Keywords: hydropower; inflow forecasting; foundation models; hydrothermal dispatch; forecast value; risk aversion; emissions

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11498205/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

When do inflow forecasts add value? Closed-loop evidence on foundation models, hedging and scarcity in the Brazilian hydrothermal system

Leonardo Forero[a,*], Harold D. de Mello Jr.[a], Antonio Guilherme Garcia Lima[a], Marco Aurélio C. Pacheco[b]

[a] Department of Electrical Engineering, Universidade do Estado do Rio de Janeiro (UERJ), Rio de Janeiro, Brazil

[b] Applied Computational Intelligence Laboratory (ICA), Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil

* Corresponding author: mendonza@ica.ele.puc-rio.br (ORCID 0009-0004-8408-1916)

Abstract

Time-series foundation models are proposed as inflow forecasters for hydropower planning, but are judged by statistical skill rather than by their effect on system operation. We asked when a better inflow forecast adds operational value, evaluating monthly forecasts for 158 Brazilian basins from 2005 to 2024 by energy error, by error weighted by the observed marginal operation cost, and in a closed-loop decision experiment: a rolling-horizon dispatch of the interconnected system using only information available at each decision. Weighted by marginal cost, Chronos-2 had R$ 37 billion less exposure than a periodic autoregressive model refitted every year and corrected for bias, an advantage that vanished once its own bias was removed. The simulator

could not reproduce thermal generation and storage at once, so the decision experiment is reported under calibrations that bound the scarcity of the system, and its simulated costs are read as comparisons between policies, not as real costs. Over the observed record, a 10% haircut and a switch to Chronos-2 both lowered regret in point estimate, but their intervals included zero and more than the whole gain came from the 2021 drought. With both models re-run on 100 hydrological histories resampled from the same years, the haircut lowered expected regret by R$ 129 billion (104 to 155) and paid in 88% of histories when the system was scarce, but in only about half when it was calibrated on storage; at equal haircut Chronos-2 was cheaper than the autoregressive model by R$ 10 to 16 billion, about a tenth of the effect of hedging. Forecasts transported from the observed record instead of re-run reversed the sign of that comparison. The native predictive distribution of Chronos-2, combined across basins with a copula, was close to calibrated one month ahead and had the lowest quantile score at every lead, a descriptive difference; its 0.25 quantile matched the best fixed haircut under scarcity, but the right quantile depended on scarcity and could not be learned before 2021. In this system forecast value is decided mostly by scarcity and by the hedging policy, and only secondarily by the choice of forecaster.

Keywords: hydropower; inflow forecasting; foundation models; hydrothermal dispatch; forecast value; risk aversion; emissions

Introduction

Brazil operates its interconnected power system (SIN) around stored water. Each month the operator decides how much water to release and how much thermal generation to dispatch, weighing the value of water in the reservoirs against the cost of fuel and the risk of a shortfall, with stochastic dual dynamic programming over equivalent reservoirs [1, 2]. Inflow forecasts are the first input of that chain, and the reference model is the periodic autoregressive family PAR(p) [3, 4].

Machine learning offers alternatives. Networks trained on many basins can forecast basins they never saw [5], and time-series foundation models such as Chronos, Chronos-Bolt and Chronos-2

forecast any series zero-shot [6, 7]. On the 158 Brazilian basins studied here, these foundation models, given their full record, reached a median Nash-Sutcliffe efficiency (NSE) [8] of 0.64 to 0.65 one month ahead against 0.61 for PAR(p) in a separate benchmark of forecasting skill by the authors (manuscript under review). Results of this kind are usually presented as evidence that a better model would improve operation.

That step from skill to value is not automatic. The value of a forecast depends on the decisions it informs and on the losses of the decision maker [9]; for hydropower, forecast skill and economic value have repeatedly been found to diverge [10–13], and energy forecasting increasingly favours decision-aware evaluation [14, 15]. In a hydrothermal system two features matter in particular. Errors are not equally costly across months, since a gigawatt-month of missing water is cheap when reservoirs are full and expensive in a drought. And operators do not use forecasts naively: risk aversion, through guide curves, conditional value-at-risk or conservative inflow scenarios, is built into planning precisely because optimistic errors are dangerous [2, 16, 17]. A forecaster can therefore look valuable because it is biased in the right direction, and the comparison between forecasters can depend on how much the operator hedges. Earlier assessments of the economic value of hydrological forecasts evaluated one forecasting system against climatology or against perfect information [10–12]; none, to our knowledge, has compared modern pretrained forecasters with a recalibrated operational model under a common hedging rule and with the uncertainty of the comparison quantified.

Our aim is not to crown a forecaster but to show when, and under which assumptions about scarcity and risk aversion, a better forecast adds operational value. We ask whether the advantage of foundation models survives three increasingly operational evaluations: energy error, absolute error weighted by the observed marginal operation cost, and a closed-loop decision experiment in which each forecast drives a dispatch policy that knows only what an operator would know at the time and faces the inflow that actually occurred. Throughout, we compare against a strong reference, PAR(p) refitted every year with all data available at the forecast origin and corrected for bias using only past errors, and we apply the same hedging rules to every model. The contributions are a reproducible pipeline from public operator data that turns basin-level inflow forecasts into energy, price-weighted exposure and simulated operation; evidence that

the foundation models' price-weighted advantage comes from the sign of their errors; and evidence that, in the closed-loop decision experiment over the observed record, both the choice of forecaster and the level of hedging turn on a single drought, so that this history cannot rank them, while the emissions cost of hedging is well determined; an evaluation of the native predictive distribution of a foundation model, aggregated to system energy and used directly for hedging; and an estimate over 100 resampled hydrological histories, with both models re-run on each, in which hedging pays in expectation when the system is scarce and the choice of forecaster, once both are hedged alike, matters about a tenth as much. We present this closed-loop decision experiment as a study of how scarcity shapes forecast value in simulated operation, judged against adequacy criteria fixed before the economic comparisons and reported under three calibrations of the simulated system, not as an estimate of what each forecast would have cost the Brazilian system.

Methods

Figure 1 summarises the pipeline. All inputs are public data from the Brazilian National Grid Operator (ONS) [18], the inflow forecasts described below and the Brazilian consumer price index (IPCA). Every quantity that enters a forecast or a policy uses only information available at the time of the forecast or the decision; observed inflows, costs and generation are used only to evaluate.

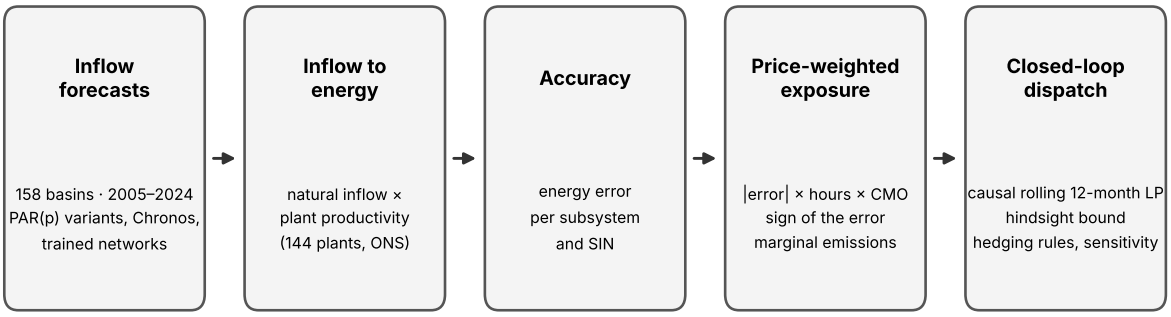


Figure 1. Pipeline of the study. Inflow forecasts are converted to natural energy inflow and evaluated by accuracy, price-weighted exposure and a closed-loop decision experiment.

Inflow forecasts

Monthly natural inflow for 158 basins of the SIN was derived from the operator's daily series for 1931 to 2024 and forecast one to twelve months ahead from every month between December 2004 and November 2024. Pooled models were trained on data up to 2004 in five folds of basins, with basins of the same cascade kept in the same fold, so that every basin was forecast by models that never saw it or its test years. The forecasts were produced for a separate benchmark of forecasting skill by the authors, currently under review; that study compares models by NSE across basins with held-out basins and years, whereas this one uses only its test-period forecasts and asks what they are worth in energy and operation, so no result is shared between the two. We used the local PAR(p), with the order of each calendar month selected by the Bayesian information criterion up to six and coefficients estimated by least squares on the monthly standardised series, fitted once on the record up to 2004 ("PAR(p) original"); the same model refitted every December on all data up to that date ("recalibrated"), so that forecasts issued in each month of a year use the parameters estimated at the end of the previous year; the local climatology refitted in the same way; pooled networks (GRU, PatchTST-S and PatchTST-M [19] pretrained on real inflow from other basins, and PatchTST-S pretrained on synthetic series that reproduce the seasonality, persistence and skewness of inflow); and the foundation models Chronos-2, Chronos-Bolt-small and Chronos-Bolt-tiny with up to 512 months of history [6, 7]. All models received the same observed record up to each forecast origin. Trained networks are averaged over five seeds; foundation-model forecasts are the median of the predictive distribution, and all forecasts are truncated at zero.

Bias correction is causal. For each model, basin and lead time, in units of flow, the forecast is reduced by the mean error over the previous 60 target months already observed at the origin (at least 12; otherwise no correction), the result is truncated at zero, and the model is labelled "debiased". Because energy is a linear combination of basin flows with fixed productivities, the correction applied per basin carries over to subsystem energy in the same way for every model. This gives every model, including PAR(p) and Chronos-2, the same opportunity to remove a systematic offset that an operator could detect.

From inflow to energy

The ONS register of reservoirs gives, for each plant, its productivity at 65% of the useful volume, its subsystem and the inflow gauge it uses; the ENA that ONS publishes for each reservoir is the natural inflow multiplied by that productivity. For the 56 plants whose daily ENA in 2000 to 2004 was strictly proportional to inflow, the factor we measured matched the register productivity with a median ratio of 1.002 and a correlation of 1.000. Linking plants to basins by gauge code, removing duplicate register entries, gave 144 plants, and the natural energy inflow of subsystem s in target month t is

$$\widehat{E}_{s,t}^{(m,h)} = \sum_{i \in s,\ i \text{ operating at } t} \rho_i \, \widehat{Q}_{i,t}^{(m,h)},$$

with ρ_i the productivity, $\widehat{Q}$ the forecast inflow of model m at lead h and the same sum with observed inflow as reference. Between 2005 and 2024 the reconstruction covered 97% of the SIN energy inflow published by ONS (98% in the North, 84% in the Northeast, 96% in the South and 102% in the Southeast/Centre-West), with a monthly correlation of 0.99. The excess in the Southeast is plausibly due to the fixed productivity at 65% of useful volume, which overstates the energy of plants operating below that level, and to differences in how the official series treats plants in cascade. For the dispatch, the reconstruction of each subsystem was rescaled by the ratio between the ONS series and our reconstruction in 2000 to 2004, a period before any test year.

Price-weighted exposure

The ONS publishes the marginal operation cost (CMO) of each subsystem for every operating week [18]. The monthly CMO is the mean of the weekly values of the operating weeks that start in the month (typically four); no month lacked a value, and 124 of the 960 subsystem-months had a CMO of R$ 1/MWh or less, when hydro plants were at the margin. We weighted the absolute energy error of each subsystem and month by that CMO,

$$X_{s,t}^{(m,h)} = \left| \widehat{E}_{s,t}^{(m,h)} - E_{s,t} \right| \cdot H_t \cdot \mathrm{CMO}_{s,t},$$

with H_t the hours of the month, deflated to December 2024 reais with the IPCA and summed over subsystems and months. An error is optimistic when $\widehat{E} > E$ (more water forecast than arrived) and pessimistic otherwise. We call X the price-weighted exposure of a forecaster. It measures how much energy a forecaster got wrong in months when energy was expensive; it is not the cost that the error caused, because the CMO was produced by the operator with its own forecasts and decisions, nor the benefit that switching forecasters would bring, which only the closed-loop decision experiment can address. We report variants that weight optimistic errors by 1.5 and 2, that cap the CMO at the 95th and 99th percentiles of each subsystem, and that rescale the error to the ONS level. Emissions were indicated by the factor of the marginal thermal plant, the plant with the highest variable unit cost (CVU) not above the CMO, with typical factors per fuel (0.45 tCO_2/MWh for natural gas, 0.75 for oil and diesel, 0.95 for coal, zero for nuclear and biomass); 98 of 691 subsystem-months with a thermal plant at the margin had an unidentified fuel, which we filled with the median of the subsystem and year and bounded with zero and the coal factor.

Closed-loop decision experiment

In the closed-loop decision experiment we simulated the operation of the SIN from January 2005 to December 2024 with one equivalent reservoir and a monthly rolling horizon. Storage capacity and the initial storage (63% in January 2005) are those published by ONS. Demand on hydro and thermal plants is their observed monthly generation from the ONS energy balance. The thermal fleet of each month is the list of plants in that month's CVU publication, with the effective capacity of operating units from the register of installed capacity at an availability of 0.85; 48% of plant-month capacities could not be matched by name and were imputed with the median of the matched ones; their distribution by year and fuel, and the share of simulated thermal generation that comes from imputed plants, are reported in the Results. Hydro generation is limited by 95% of the installed hydro capacity, and unserved energy costs R$ 5,000/MWh at December 2024 prices.

At the start of month t the policy knows the inflow forecast issued at the end of month $t-1$, realised demand, inflow and generation up to $t-1$, and the CVU list of month $t-1$. Demand for months t to $t+11$ is forecast as the demand of the same calendar month in the last observed year multiplied by the growth of the last twelve months over the twelve before, raised to the number of years ahead; its mean absolute error was 3.0% one month ahead and 3.8% twelve months ahead, with a bias below 0.5%. The thermal fleet, the hydro limit and the storage capacity known at the decision are held fixed over the horizon; the capacity of the fleet known at the decision differed from that of the month by 2.7% on average. Only the execution of month t uses the realised demand, fleet and inflow. Table 1 lists every input, when it becomes available and where it is used. An earlier version of this study planned with the realised demand and fleet of the whole horizon; we report that protocol for comparison.

With $k = 1, \ldots, K$ indexing the months of the horizon ($K = 12$, shorter at the end of the period), b the plants of the known fleet with capacity $\bar{g}_b$ and cost c_b, $\hat{A}_k$ the inflow forecast, $\hat{d}_k$ the demand forecast, $\bar{S}$ and $\bar{H}$ the storage and hydro limits and G_K the guide-curve storage for the last month, the policy solves

$$\min \sum_{k=1}^{K} 730 \Big(\sum_{b} c_b \, g_{b,k} + c^{\mathrm{def}} D_k \Big) + 730 \, \pi \, u$$

subject to $\sum_b g_{b,k} + H_k + D_k = \hat{d}_k$, $S_k = S_{k-1} + \hat{A}_k - H_k - P_k$, $0 \le g_{b,k} \le \bar{g}_b$, $0 \le H_k \le \bar{H}$, $0 \le S_k \le \bar{S}$, $D_k, P_k \ge 0$, $S_K + u \ge G_K$ and $u \ge 0$. Energies are in average megawatts over the month (730 hours), S_0 is the current storage, P_k is spillage, π is the highest CVU of the fleet and there is no discounting. The guide curve is the median stored-energy fraction of each calendar month in 2000 to 2004 times the storage capacity. Execution sets thermal generation to the planned value of the first month, capped by the available capacity and the demand, generates hydro power with the realised inflow up to the demand, the stored water and the hydro limit, meets any remaining shortfall with the unused thermal capacity and then with deficit, and spills storage above capacity.

Operating cost in the experiment is the fuel cost of thermal generation, dispatched in merit order of CVU, plus deficit times the deficit cost, deflated month by month to December 2024. The guide-curve penalty enters only the planning problem and is not counted as cost. Every cost

and regret reported for the experiment is a simulated cost of this simplified system in December 2024 reais, used to compare policies under the same assumptions; none is an estimate of the cost incurred by the real system or of the economic value of a forecast to consumers. Emissions are thermal generation times the factor of each plant's fuel.

Table 1. Information used in the closed-loop decision experiment at the decision for month t.

Input	Available	Source	Planning (months t to t+11)	Execution (month t)
Inflow forecast	end of $t-1$	model under evaluation	forecast for t to t+11	not used
Observed inflow	after each month	ONS	up to $t-1$ (quantiles, bias, nested rule)	realised inflow of t
Demand	after each month	ONS energy balance	forecast from data up to $t-1$	realised demand of t
Thermal fleet and CVU	start of $t-1$	ONS CVU list and capacity register	list of $t-1$, held fixed	list of t
Hydro limit, storage capacity	start of t	ONS register and EAR	value of t, held fixed	value of t
Guide curve	before 2005	ONS stored energy, 2000 to 2004	terminal target	not used
Deficit cost, IPCA	fixed, monthly	assumption, Central Bank	cost coefficients	cost accounting

Two references anchor the comparison. The perfect-inflow policy is the same rolling-horizon rule fed with the observed inflow of the next twelve months and with the same demand forecast and fleet; regret is the cost of a policy minus the cost of this policy, and measures the value of inflow information inside this decision rule; it is not a universal measure of the economic value of a forecast, and every regret in the paper is relative to this rule. It is not a lower bound: the rule looks only twelve months ahead and values end-of-horizon water only through the guide-curve penalty, so a pessimistic forecast can act as an implicit risk premium and beat it. We therefore also solved the hindsight optimum, a single linear programme over the 240 months with the realised inflow, demand and fleet and the same terminal condition, which is a lower bound on the cost of any policy in the simplified system. Because policies end the period with different storage, we also report regret with the difference in final storage from the perfect-inflow policy valued at the median CVU of December 2024 (R$ 401/MWh).

Because an equivalent reservoir uses water more freely than the real system, demand was scaled by a factor that sets how scarce the simulated system is. Before comparing any policies we fixed

five adequacy criteria for operation with the original PAR(p), the model in operational use during the period: mean thermal generation within 15% of the observed value, mean stored energy within 10 percentage points, a monthly correlation of simulated and observed storage of at least 0.8, a number of months with storage below 30% between half and twice the observed number, and no more than three months with deficit. We report results under three calibrations of the factor: on mean thermal generation, the main case of the earlier version; on mean stored energy; and a compromise that minimises the sum of the squared deviations of thermal generation and storage, each divided by its tolerance, over a grid from 0.95 to 1.25. No single calibration meets every criterion, so the economic comparisons are reported under all three and read as conditional on the scarcity each one implies. We also compared annual simulated thermal generation, storage, cost and marginal cost with the observed operation.

Hedging was applied identically to every model in two ways that we report separately. A fixed haircut of 2.5% to 25% multiplies every forecast by the same factor, so the adjustment is identical across models. An empirical quantile multiplies the forecast by the 10%, 25% or 50% quantile of the ratio between observed and forecast SIN energy over the previous 60 target months (at least 12; ratios with a non-positive forecast are discarded and the result is truncated at zero); the rule is common, but the adjustment uses each model's own error distribution and therefore rewards models whose errors are smaller. To check whether the haircut could have been chosen without hindsight, we also ran a nested rule that, at the start of each year, adopts the haircut of the grid with the lowest simulated cost over the previous years, with no haircut until three (or five) years are available.

Native predictive quantiles

Fixed haircuts and empirical quantiles adjust a point forecast from outside the model. Chronos-2 also produces a predictive distribution, which we evaluated directly. For every basin and forecast origin we stored its 21 predictive quantiles (0.01 to 0.99) with the same long context as the median. Quantiles of the SIN energy cannot be obtained by adding basin quantiles unless all basins move together, so we combined the basin distributions with a Gaussian copula: the dependence between basins is the correlation of their standardised monthly log-inflow anomalies

from 1931 to 2004, before any test year; 2,000 correlated draws per origin and lead are mapped through each basin's quantile function, weighted by plant productivity and summed. We report the comonotonic sum as an upper bound on the spread. The calibration of these distributions, which we compare descriptively without a formal test, was measured on the SIN energy for leads of 1, 3, 6 and 12 months by the coverage of the central 50%, 80% and 90% intervals, the observed frequency below each quantile and the mean quantile score over 13 levels, a discrete approximation of the continuous ranked probability score, and compared with empirical quantiles of the debiased and uncorrected PAR(p) and Chronos-2. In the closed loop, a policy plans with a fixed native quantile between 0.1 and 0.5, or with the quantile chosen at the start of each year as the one with the lowest simulated cost over the previous years (median until three years are available), the probabilistic analogue of the nested haircut.

Resampled hydrological histories

The observed period contains one sequence of droughts, and a few months of it can decide a comparison. To estimate expected performance, we built 100 alternative twenty-year histories by drawing blocks of three consecutive hydrological years from 2005 to 2024 with replacement [20], which keeps multi-year droughts together, while demand, the thermal fleet, costs and deflators stay in calendar order. In each history the inflow of each basin in a month is that of the same calendar month of its source year, and the energy inflow of the SIN is computed from these flows with the plants operating in that calendar month, so that the system receives the energy its installed plants would have produced. We used two kinds of forecasts in these histories. In the transported forecasts, the forecast issued in a month keeps, for every lead, the ratio between forecast and observed inflow that the model achieved from the source month, applied to the inflow of the target month in the new history; within a block the forecast is identical to the original, and across block boundaries it carries the relative error the model made at that lead. This preserves an estimated error structure and is a conditional sensitivity analysis, not a counterfactual forecast, because a model given a different history would have produced different forecasts. We therefore also re-ran the main models on every synthetic history: the recalibrated PAR(p) was refitted each December on the synthetic record up to that date, and Chronos-2 was run zero-shot with up to

512 months of synthetic history, with the observed record before 2005 as the common past; both were debiased with the same causal rule and converted to energy with the target-month plant configuration. The native 0.25 quantile of the re-run Chronos-2 was aggregated with the same copula. The original sequence was run as an additional history: the re-run PAR(p) reproduced the forecasts of the observed record exactly and the re-run Chronos-2 differed from them by 0.11% on average and 0.9% at most, a difference of numerical precision on the GPU. In every history, under the three calibrations, we ran the debiased PAR(p) and Chronos-2 with and without a 10% haircut, the uncorrected Chronos-2 and its native 0.25 quantile, each with re-run and with transported forecasts. For each policy we report the mean regret with a 95% bootstrap interval over histories, its 5th and 95th percentiles and its 90% conditional value at risk (the mean of the worst 10% of histories), and for pairs of policies the share of histories in which one is cheaper. The interval of a mean narrows as histories are added; the share of histories is the quantity comparable with the uncertainty of a single record. Because the histories reuse the observed years, they change the order and frequency of the droughts of 2014, 2015 and 2021 but contain none more severe, and they describe variation over reorderings of the observed record rather than uncertainty about future climate.

Statistics

Differences between two policies or forecasters are computed month by month and summed over the 240 months. Intervals come from a moving-block bootstrap with 12-month blocks and 2,000 replicates [21], in which both members of a pair are resampled on the same blocks; we checked blocks of 6 and 24 months. Two-sided bootstrap p values are corrected with Holm's procedure [22] within each family of comparisons: for the price-weighted exposure, all models against the same reference; in closed loop, models without hedging, models at the same haircut, models at the same empirical quantile, and hedging levels for a given model. Because a non-significant difference does not establish equivalence, we report the 90% interval and the smallest margin that contains it, and test equivalence with two one-sided tests [23] at a margin fixed in advance at 5% of the cost of the perfect-inflow policy (R$ 47 billion). We also report the share of each difference that comes from 2021, the difference without it, and the number of years in which one policy

was cheaper than the other. Bootstrap replicates use a fixed seed, and the simulator configuration, the imputation rules for the thermal fleet and the seeds are recorded in the repository together with the notebooks that rebuild every table and figure.

The software was developed with the assistance of a large language model (Claude, Anthropic) for code writing and data processing, and the same tool was used for language editing; every model, experiment and analysis was specified by the authors, who checked all code, results and text and take full responsibility for them.

Results

Accuracy

Chronos-Bolt-tiny and Chronos-2 had the highest median NSE across basins one month ahead (0.648 and 0.643), followed by the recalibrated and debiased PAR(p) (0.636); the original PAR(p) reached 0.611. In energy terms the order changed (Table 2). One month ahead, the recalibrated and debiased PAR(p) had the lowest mean absolute error of the SIN energy inflow (8.52 GWmed, against 8.83 for Chronos-2 and 8.88 for the original PAR(p)). At three months the recalibrated PAR(p) and the foundation models were level (11.36 to 11.48 GWmed), and at twelve months the foundation models were best (12.81 against 13.08). The models differed most in the sign of their errors. The original PAR(p) over-forecast, increasingly with lead (+1.6 GWmed at one month and +8.3 at twelve), whereas Chronos-2 under-forecast at one month (-2.8 GWmed).

Table 2. Accuracy of the SIN energy inflow forecasts, 2005 to 2024 (GWmed).

Model	MAE 1 month	Bias 1 month	MAE 3 months	MAE 12 months	Bias 12 months
PAR(p) recalibrated, debiased	8.52	+0.24	11.39	13.08	+4.03
PAR(p) recalibrated	8.72	+1.42	11.64	14.48	+7.88
PAR(p) original	8.88	+1.62	11.87	14.80	+8.27
Chronos-2	8.83	−2.85	11.39	12.81	+0.70
Chronos-2, debiased	8.84	−0.28	11.60	13.39	+2.22
Chronos-Bolt-small	8.83	−1.78	11.48	12.96	+2.31
Chronos-Bolt-tiny	8.92	−2.20	11.36	12.81	+1.23
PatchTST-S (synthetic pretraining)	9.09	+1.46	12.51	14.37	+6.20
PatchTST-M	9.32	+0.83	12.82	14.31	+4.92
GRU	9.26	+1.23	12.88	14.14	+5.99
Climatology, recalibrated	11.95	+5.11	12.76	14.77	+8.28

Price-weighted exposure

Weighted by the CMO (Figure 2), Chronos-2 had R$ 37 billion less exposure than the recalibrated and debiased PAR(p) over twenty years one month ahead (95% interval 13 to 70; Holm-adjusted p = 0.03 among the 17 models compared), and less exposure in 14 of 20 years. The difference held when any single year was left out (R$ 29 to 38 billion), without the drought years (R$ 21 billion), with blocks of 6 and 24 months, with the CMO capped at its 99th or 95th percentile (R$ 33 billion in both cases, intervals 10 to 63 and 14 to 62) and when the error was rescaled to the ONS level (R$ 41 billion). Chronos-Bolt-tiny and Chronos-Bolt-small had R$ 22 and 21 billion less exposure, with intervals reaching zero. None of the trained networks did better than PAR(p).

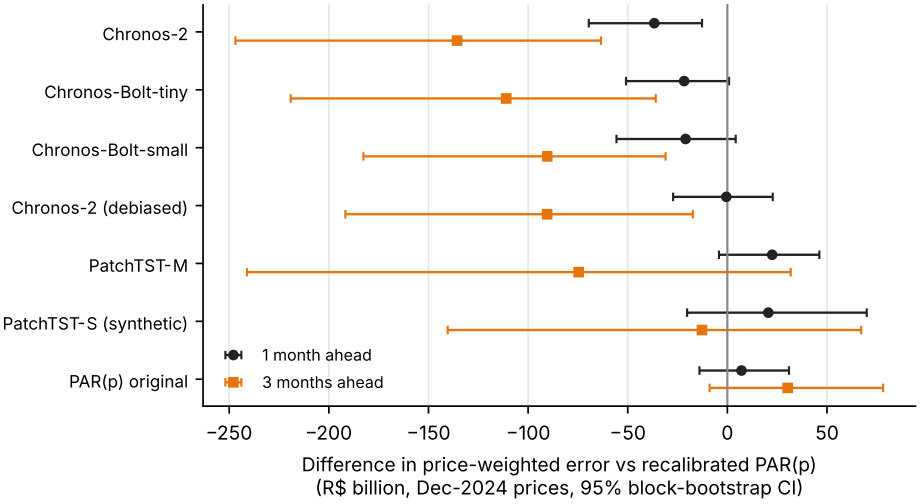


Figure 2. Price-weighted exposure relative to the recalibrated and debiased PAR(p). Differences over 2005 to 2024 one and three months ahead, in R$ billion at December 2024 prices, with 95% block-bootstrap intervals. Negative values favour the model.

The advantage of Chronos-2 at one month disappeared once its own bias was removed (−R$ 0.5 billion, interval −27 to 23; between −2.9 and +7.8 under every CMO variant). It comes from the direction of its errors (Figure 3). The recalibrated PAR(p) accumulated R$ 367 billion of exposure in months in which it forecast more water than arrived and R$ 230 billion in the other months; for Chronos-2 the split was R$ 287 and 274 billion. Optimistic errors fall disproportionately in dry, expensive months, and a forecaster that leans towards less water makes fewer of them. Weighting optimistic errors more heavily widened the gap to R$ 77 billion with a weight of 1.5 and R$ 117 billion with 2, and left the debiased Chronos-2 level with PAR(p).

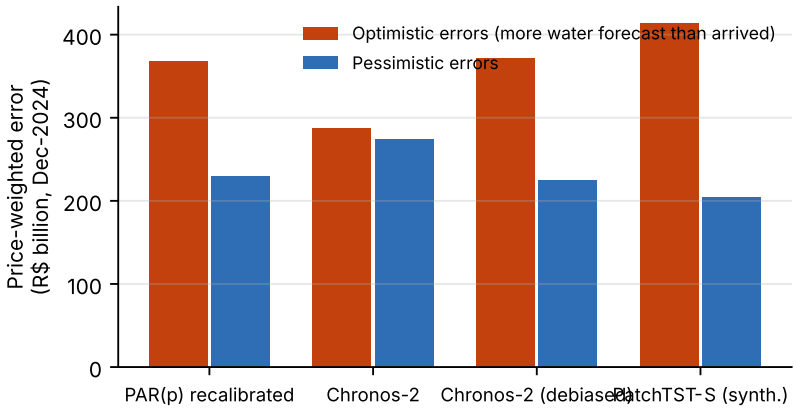


Figure 3. Price-weighted exposure one month ahead by sign of the error. Optimistic errors forecast more water than arrived.

Three months ahead the gap was larger and partly independent of bias. Chronos-2 had R$ 136 billion less exposure than the recalibrated and debiased PAR(p) (63 to 247; Holm-adjusted $p < 0.001$), and its debiased version still R$ 90 billion less (17 to 192; unadjusted p between 0.002 and 0.015 depending on block length, Holm-adjusted $p = 0.05$ among the 17 models), R$ 81 billion with the CMO capped at the 99th percentile and R$ 60 billion at the 95th. The emission indicator followed the same pattern: 42 $MtCO_2$ lower for Chronos-2 at one month (3 to 91) and 102 $MtCO_2$ at three months (47 to 173). These are differences in exposure. Whether they translate into lower cost depends on how the forecasts are used, which the closed loop addresses.

Adequacy of the simulator

Table 3 checks the simulated operation with the original PAR(p) against the five criteria fixed in the Methods, and no calibration met more than two of them. With demand scaled by 1.196, the simulator reproduced the observed mean thermal generation of 9.0 GWmed; monthly simulated and observed thermal generation correlated at 0.79 and stored energy at 0.88, annual simulated cost and the cost of the observed thermal generation dispatched in merit order on the same fleet at 0.47, and the simulated marginal cost and the observed CMO at 0.66 (Spearman). Mean storage, however, was 15 percentage points below the observed, storage was below 30% in 93 months against 27 observed, and 13 months had deficit. The simulator reproduced the dry years 2014 and 2015 closely (thermal 14.9 and 13.8 GWmed against 15.6 and 15.5 observed) but was too dry from 2016 onwards: simulated storage fell to almost nothing between 2017 and 2021,

whereas observed storage stayed between 18% and 34%, and the simulated marginal cost of 2021 reached R$ 2,484/MWh against an observed mean CMO of R$ 534/MWh. Calibrating on mean storage (factor 1.02) matched the observed storage but produced 1.9 GWmed of thermal generation, a fifth of the observed.

The scan of the demand factor shows why no compromise exists in this representation: thermal generation within 15% of the observed requires a factor of at least 1.17, while storage within 10 points of the observed requires a factor of at most 1.13. The compromise calibration landed at 1.18, next to the thermal calibration, and led to the same conclusions (Table 6). We therefore read the thermal and storage calibrations as bounds on scarcity rather than as competing estimates of the real system: the first exaggerates the scarcity of the late 2010s and of 2021, the second removes nearly all of it, and the operation of the SIN lies between them. Neither supports reading the simulated costs as operational costs; they are used only to compare policies under a given degree of scarcity.

Table 3. Adequacy criteria for operation with the original PAR(p) under three calibrations of the demand factor.

Criterion	Required	Thermal (1.20)	Storage (1.02)	Compromise (1.18)
Mean thermal generation, simulated / observed	0.85 to 1.15	1.00, met	0.21, not met	0.92, met
Mean stored energy, bias (percentage points)	−10 to +10	−14.5, not met	−0.1, met	−13.0, not met
Monthly correlation of stored energy	at least 0.80	0.88, met	0.84, met	0.87, met
Months with storage below 30% (27 observed)	14 to 54	93, not met	58, not met	91, not met
Months with deficit	at most 3	13, not met	6, not met	13, not met
Criteria met		2 of 5	2 of 5	2 of 5
Monthly correlation of thermal generation		0.79	0.62	0.79
Marginal cost and CMO, Spearman correlation		0.66	0.57	0.67

The criteria were fixed before the economic comparisons. The full scan of the factor from 0.95 to 1.25 is in Additional file 1.

The thermal capacity imputed for unmatched plants was not concentrated in the critical years. Imputed plants held 40% of the available thermal capacity on average, a share that fell from 50% to 55% in 2008 to 2011 to between 27% and 36% from 2019 onwards; in 2021 it was

34%, and plants with imputed capacity supplied 34% of the simulated thermal generation of that year and 34% of the observed generation dispatched in merit order. Imputation fell mostly on plants whose fuel was also missing from the register (2.7 GW on average, all imputed), gas (44% of plant-months) and coal (50%), while oil plants were almost all matched (7%). The distribution by year and fuel is in Additional file 1; the effect of the imputed capacity on the results is bounded by the sensitivity analysis (Figure 6).

Decision experiment over the observed history

Table 4 summarises twenty years of simulated operation with information available at each decision. All regrets are relative to the perfect-inflow rule; as a share of its cost, R$ 943 billion, the regret of the best hedged policies was 7% to 9% and that of the recalibrated and debiased PAR(p) 33%. The hindsight optimum cost R$ 765 billion and the perfect-inflow policy R$ 943 billion; the gap of R$ 178 billion is the cost of the rolling-horizon rule itself, before any forecast error. Among point forecasts, Chronos-2 had the lowest regret (R$ 190 billion), ahead of the Chronos-Bolt models (223 and 228), the debiased Chronos-2 (237), the recalibrated and debiased PAR(p) (308), PatchTST-M (316), the original PAR(p) (619) and the climatology (814). Hedged policies did better than any point forecast. The hindsight optimum kept storage well above every policy from 2012 to 2020 and drew it down only in 2021 (Figure 5), which is the behaviour that hedging tries to approximate without knowing when the drought will come. The ordering of unhedged forecasts was the same under the earlier protocol, with regrets that differed by less than R$ 60 billion except for the climatology (96); hedged policies gained more from planning with forecast demand, the debiased Chronos-2 with a 10% haircut falling from 128 to 71. Valuing the difference in final storage changed regrets by less than R$ 21 billion.

Table 4. Closed-loop operation from 2005 to 2024 with information available at each decision. Calibrated system; costs in R$ billion at December 2024 prices; regret is relative to the perfect-inflow policy.

Policy	Operating cost	Regret	Above hindsight optimum	Deficit (TWh)	Thermal (TWh)	CO_2 (Mt)	Months below 20%
Hindsight optimum	765	−178	0	0.0	1778	1061	6
Perfect inflow, rolling horizon	943	0	178	16.1	1767	1064	6

Policy	Operating cost	Regret	Above hindsight optimum	Deficit (TWh)	Thermal (TWh)	CO_2 (Mt)	Months below 20%
PAR(p) recalibrated, debiased, −15%	1008	65	243	0.0	1946	1178	4
PAR(p) recalibrated, empirical q0.25	1014	71	249	0.0	1908	1160	4
Chronos-2, debiased, −10%	1014	71	249	3.9	1895	1146	5
Chronos-Bolt-tiny, debiased, −10%	1026	83	261	6.8	1883	1138	5
Chronos-2, empirical q0.25	1029	87	265	0.0	1922	1168	3
PAR(p) recalibrated, debiased, −10%	1066	124	302	20.8	1869	1128	5
Chronos-2	1132	190	368	40.4	1801	1088	16
Chronos-Bolt-tiny	1165	223	401	50.2	1769	1069	17
Chronos-2, debiased	1180	237	415	51.4	1748	1059	35
PAR(p) recalibrated, debiased	1251	308	486	72.8	1723	1040	53
PatchTST-M	1258	316	494	69.4	1694	1028	51
PAR(p) original	1561	619	797	166.5	1571	940	58
Climatology, recalibrated	1756	814	992	219.5	1520	905	69

None of these differences between forecasters was statistically clear (Table 5). Chronos-2 was R$ 118 billion cheaper than the recalibrated and debiased PAR(p), with an interval from −484 to +85; the interval is wide and asymmetric because the difference is a sum of monthly differences that are close to zero in most months and large in a few, and 122% of it came from 2021, without which Chronos-2 was R$ 26 billion more expensive. The same held for the debiased foundation models and for every equal haircut up to 10%: the smallest margin within which the 90% interval fell was R$ 190 to 270 billion, far beyond any operationally negligible difference. Equivalence within the R$ 47 billion margin was established only where hedging was strong enough to remove deficit, at a 15% haircut for Chronos-2 (+R$ 21 billion, 90% interval −15 to 44) and at the 10% empirical quantile, where all policies converge. With empirical quantiles that use each model's own errors the foundation models were not better either (+R$ 15 billion at the 25% quantile, −31 to 56). Results with blocks of 6 and 24 months were the same in substance.

Table 5. Paired differences in operating cost over the observed history. Differences a − b in R$ billion, with 95% block-bootstrap intervals, Holm-adjusted p values within each family, the share of the difference that comes from 2021, the difference without 2021, the number of years in which a was cheaper than b, and the difference in emissions.

a	b	Difference	95% interval	Holm p	Share from 2021	Without 2021	Years a cheaper	CO_2 (Mt)
Chronos-2	PAR(p) rec., debiased	−118	−484 to 85	1.00	122%	+26	6 of 20	+47 (6 to 92)
Chronos-2, debiased	PAR(p) rec., debiased	−71	−303 to 47	1.00	125%	+18	7 of 20	+19 (−18 to 38)
PAR(p) original	PAR(p) rec., debiased	+310	−81 to 793	0.74	15%	+263	10 of 20	−100 (−187 to −42)
Chronos-2, deb., −10%	PAR(p), deb., −10%	−52	−253 to 47	1.00	163%	+33	6 of 20	+18 (−18 to 36)
Chronos-2, deb., −15%	PAR(p), deb., −15%	+21	−20 to 49	1.00	0%	+21	7 of 20	+8 (−31 to 29)
PAR(p), deb., −10%	PAR(p), deb.	−184	−772 to 120	1.00	131%	+57	4 of 20	+88 (50 to 141)
PAR(p), deb., −15%	PAR(p), deb.	−243	−1041 to 181	1.00	143%	+104	4 of 20	+138 (75 to 221)
Chronos-2, deb., −10%	Chronos-2, deb.	−166	−706 to 118	1.00	144%	+72	3 of 20	+87 (45 to 143)

Hedging as insurance

Hedging changed regret more than any change of forecaster in the point estimates, but with the same fragility. Lowering the recalibrated PAR(p) forecast cut its regret from R$ 308 billion to 124 with 10% and 65 with 15%, after which further hedging raised it again (135 at 20% and 215 at 25%); for the debiased Chronos-2 the minimum, about R$ 70 billion, was reached at 10% to 12.5% (Figure 4). The gain of a 10% haircut on PAR(p), R$ 184 billion, had an interval from −772 to +120: 131% of it came from 2021, and without that year the haircut cost R$ 57 billion more than no haircut. Its price, in contrast, was paid every year and was well determined: 88 $MtCO_2$ (50 to 141) and 147 TWh of additional thermal generation over twenty years, in exchange for 52 TWh less deficit. Along the frontier each 2.5 points of haircut added 20 to 26 $MtCO_2$ until deficit disappeared, at 12.5% to 15%, and beyond that point hedging bought nothing but fuel. These are indicators of operational emissions from the simulated thermal dispatch, computed with typical factors per fuel, and not a life-cycle assessment or an estimate of the incremental emissions of the real system. The difference between policies was insensitive to the factors used: with imputed fuels set to gas or coal, or every factor at the low or high end of its

typical range, the extra emissions of a 10% haircut were between 74 and 102 Mt.

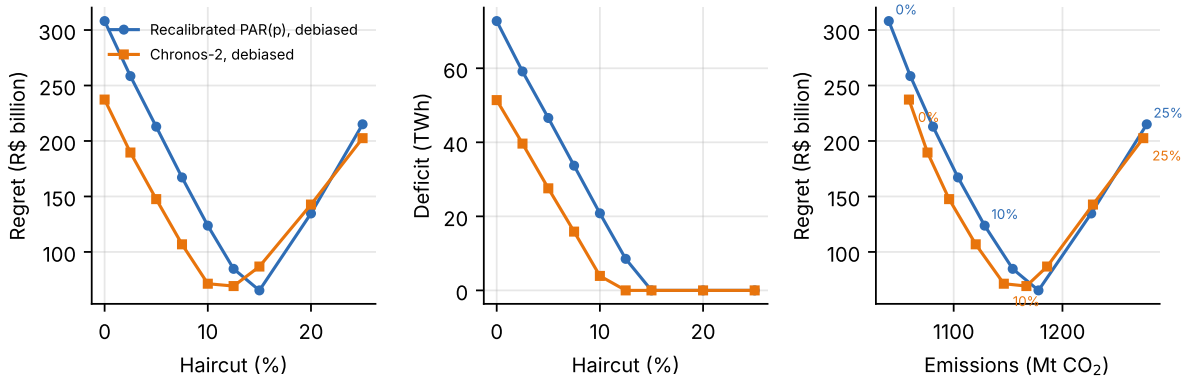


Figure 4. Regret, deficit and emissions as the haircut on the inflow forecast increases. Recalibrated and debiased PAR(p) and debiased Chronos-2 over the observed history, with information available at each decision; regret is relative to the perfect-inflow policy.

The level of hedging could not have been chosen in advance. The nested rule, which at the start of each year adopts the haircut with the lowest simulated cost over the previous years, kept PAR(p) without haircut until 2017, adopted 2.5% from 2018 and 15% only from 2022, after the drought; for Chronos-2 it adopted 10% from 2022. Its cost was within R$ 6 billion of no hedging for both models (-5.7, interval -52 to 74, and +0.8, -16 to 47), and R$ 167 to 179 billion above the fixed 10% haircut chosen with hindsight. In the observed record the years before 2021 offered no evidence that hedging would pay. The fixed haircuts of Figure 4 are therefore an ex post description of what hedging could have achieved, not an implementable policy; the nested rule is the closer analogue of what an operator could have done.

Results under the three calibrations

Because the calibration on thermal generation makes the late 2010s drier than observed, we report the main comparisons with equal weight under the storage and compromise calibrations (Table 6). The compromise reproduced the thermal case: a 10% haircut lowered the cost of PAR(p) by R$ 170 billion, with an interval from -697 to +113, and at equal haircut the two forecasters differed by R$ 28 billion (-15 to 59). Under the storage calibration the system was slack: unhedged regrets were three to six times smaller than under the thermal calibration, none of the recalibrated or foundation-model policies ran a deficit, the recalibrated and debiased PAR(p) had the lowest regret of the four, a 10% haircut raised its cost by R$ 18 billion (6 to 34)

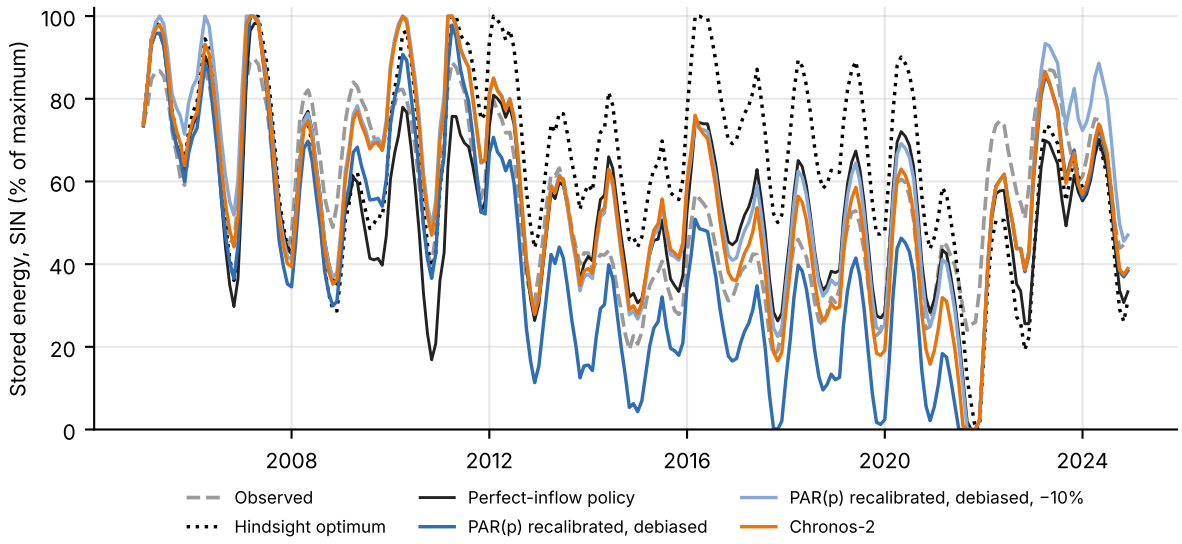


Figure 5. Stored energy of the SIN, observed and simulated. Simulations with the hindsight optimum, the perfect-inflow policy, the recalibrated and debiased PAR(p) with and without a 10% haircut, and Chronos-2.

and its emissions by 48 Mt, and the debiased Chronos-2 was more expensive than PAR(p) with or without hedging, by R$ 12 billion (0 to 28) and 20 billion (7 to 37). Under both calibrations the foundation model showed no clear advantage once its bias was removed; what changed was whether hedging paid, which depends on how scarce the system is.

Table 6. Main comparisons of the decision experiment under three calibrations of the demand factor. Regret relative to the perfect-inflow policy and differences a − b in R$ billion, with 95% block-bootstrap intervals.

Quantity	Thermal (1.20)	Storage (1.02)	Compromise (1.18)
Regret, PAR(p) recalibrated, debiased	308	54	292
Regret, PAR(p) recalibrated, debiased, −10%	124	72	122
Regret, Chronos-2, debiased	237	66	216
Regret, Chronos-2, debiased, −10%	71	92	150
Chronos-2 deb. − PAR(p) deb., no haircut	−71 (−303 to 47)	+12 (0 to 28)	−76 (−325 to 44)
Chronos-2 deb. − PAR(p) deb., both −10%	−52 (−253 to 47)	+20 (7 to 37)	+28 (−15 to 59)
Effect of a 10% haircut on PAR(p)	−184 (−772 to 120)	+18 (6 to 34)	−170 (−697 to 113)
Extra emissions of a 10% haircut on PAR(p), Mt	+88	+48	+85

Sensitivity to the system representation

How much hedging was worth depended on how scarce the simulated system was, while the comparison between forecasters at equal hedging changed little (Figure 6). In the variants with

less scarcity (demand factors of 1.00 and 1.10, the factor calibrated on storage, and imputed thermal capacity at the third quartile of matched plants) a 10% haircut raised the cost of PAR(p) by R$ 13 to 54 billion, with every interval above zero, and added 42 to 84 $MtCO_2$. In the variants with more scarcity (thermal availability of 0.75, imputed capacity at the first quartile, or plants with imputed capacity removed) the haircut lowered cost by R$ 150 to 181 billion, with intervals that reached zero in all but the last, the most extreme case. At equal 10% haircut, the debiased Chronos-2 was more expensive than the debiased PAR(p) in eight of the ten variants, by R$ 19 to 37 billion, significantly so in three of them, and cheaper only in the calibrated system and with availability of 0.75, without significance.

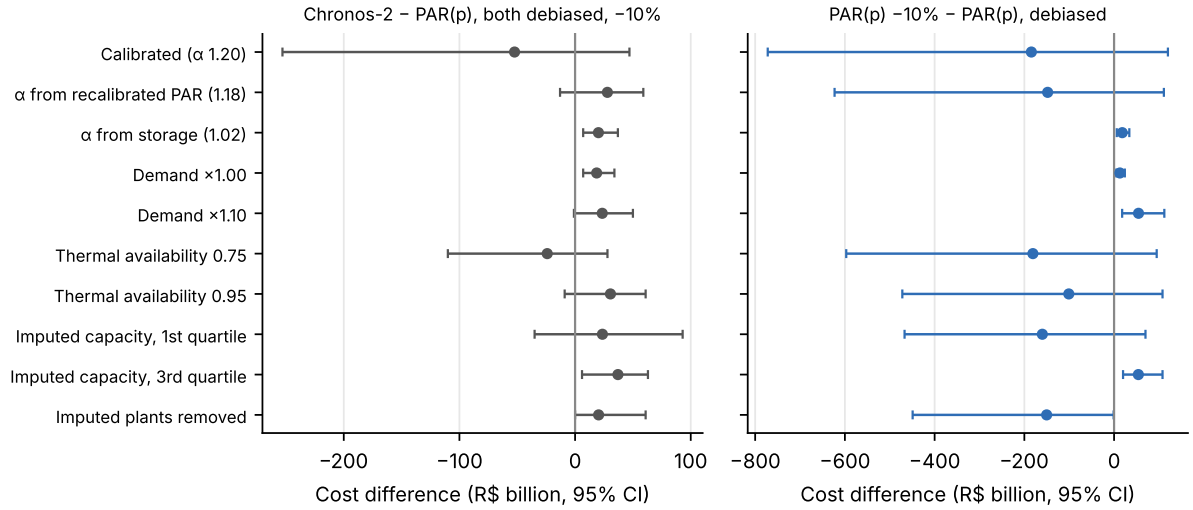


Figure 6. Forecaster and hedging effects under alternative system representations. Left: cost difference between the debiased Chronos-2 and the debiased PAR(p) at the same 10% haircut. Right: effect of a 10% haircut on the debiased PAR(p). Intervals are 95% block-bootstrap intervals.

Native predictive quantiles

The median of the stored predictive distribution reproduced the Chronos-2 forecasts used in the rest of the paper (median relative difference 0.15%). Combined across basins with the copula, the native distribution of the SIN energy was close to calibrated one month ahead: the central 80% interval covered 80% of the observations and the 90% interval 88%, and 12% and 29% of observations fell below the 0.1 and 0.25 quantiles (Table 7, Figure 7). At longer leads it was somewhat too narrow, with 72% to 74% coverage of the 80% interval. The comonotonic sum was much too wide (94% to 95%) and too low, with 1% to 2% of observations below its 0.1

quantile, because it assumes that all basins run dry together. Empirical quantiles of PAR(p) and Chronos-2 under-covered at every lead, 62% to 72% for the 80% interval. The native distribution had the lowest quantile score at every lead, 5.5% of mean energy one month ahead against 5.6% for the empirical quantiles of PAR(p) and 5.8% for those of Chronos-2, and 7.0% against 7.4% and 7.7% twelve months ahead. These differences are descriptive: we did not test them formally, and with 240 serially correlated months per lead the smaller ones, such as 5.5% against 5.6% one month ahead, should not be read as an established advantage. In the dry years 2014, 2015 and 2021 its 80% interval covered between 67% and 83% of the observations.

Table 7. Calibration of predictive distributions of the SIN natural energy inflow, 2005 to 2024.

Distribution	Lead	80% cover	90% cover	Below q0.10	Below q0.25	Score (%)
Chronos-2 native, copula	1	0.80	0.88	0.12	0.29	5.52
Chronos-2 native, copula	6	0.72	0.87	0.17	0.34	6.92
Chronos-2 native, copula	12	0.73	0.84	0.19	0.33	6.99
Chronos-2 native, comonotonic	1	0.94	0.99	0.02	0.14	5.93
Chronos-2 native, comonotonic	12	0.95	0.99	0.02	0.16	7.31
PAR(p) recalibrated, empirical	1	0.72	0.81	0.14	0.27	5.61
PAR(p) recalibrated, empirical	6	0.67	0.79	0.19	0.34	7.21
PAR(p) recalibrated, empirical	12	0.65	0.74	0.23	0.35	7.42
Chronos-2, empirical	1	0.72	0.80	0.15	0.29	5.78
Chronos-2, empirical	6	0.69	0.77	0.16	0.30	7.47
Chronos-2, empirical	12	0.63	0.75	0.19	0.31	7.72

Lead in months. A calibrated distribution has coverages of 0.80 and 0.90 and frequencies of 0.10 and 0.25 below the quantiles. Score: mean pinball loss over 13 levels from 0.05 to 0.95, in % of mean energy inflow.

In the decision experiment the native quantiles behaved like a calibrated haircut (Table 8). Under the thermal calibration, planning with the native 0.25 quantile gave a regret of R$ 71 billion with no deficit, the same as the debiased Chronos-2 with a 10% haircut (difference -0.1, interval -66 to 74) and the empirical 0.25 quantile of PAR(p); the 0.2 and 0.3 quantiles cost more (111 and 91). Under the storage calibration the native median was the cheapest policy (R$ 51 billion), cheaper than the median Chronos-2 forecast by R$ 9 billion (4 to 16) and close to the debiased PAR(p) (54), while the 0.25 quantile cost R$ 46 billion more than that PAR(p) (16 to 84). Under the compromise the native 0.3 quantile (138) came close to the cheapest policy, PAR(p) with a 10% haircut (122). The native distribution was a better basis for hedging than

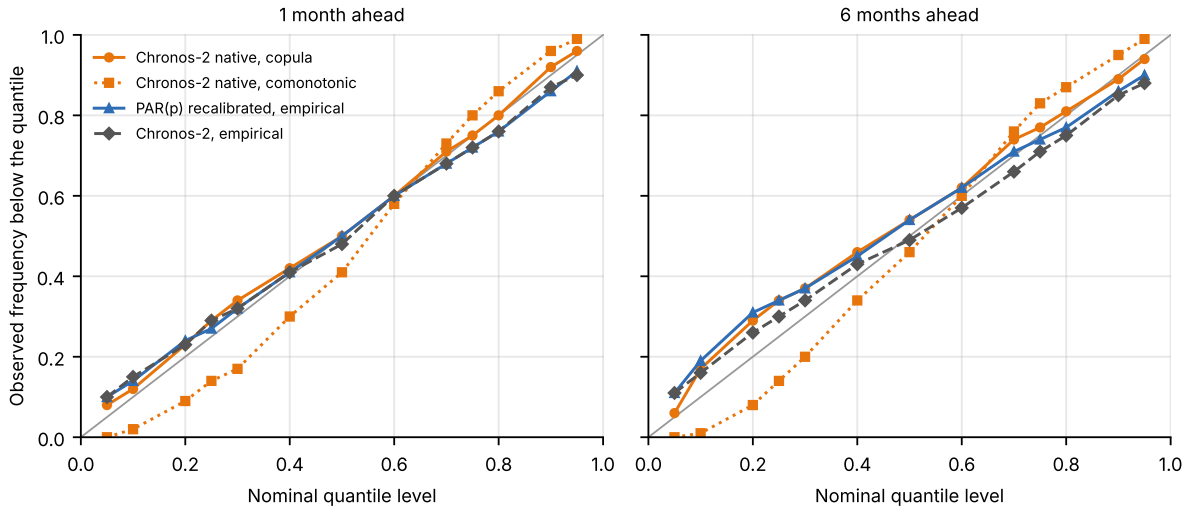


Figure 7. Reliability of predictive quantiles of the SIN natural energy inflow, one and six months ahead. A calibrated distribution follows the diagonal.

empirical quantiles of the same model: at the 0.25 quantile it was cheaper by R$ 15 billion under the thermal calibration (−71 to 9), 64 billion under the storage calibration (33 to 113) and 49 billion under the compromise (17 to 106), and the comonotonic 0.25 quantile over-hedged at R$ 142 to 221 billion more. The quantile chosen each year from past costs behaved like the nested haircut: it stayed at the median until 2021 and moved to 0.25 (0.3 under the compromise) only from 2022, and cost about as much as the fixed median. A calibrated predictive distribution lets an operator express caution without an arbitrary haircut, but the right quantile still depends on how scarce the system is, and the record before 2021 gave no signal for choosing it.

Table 8. Regret of policies using native quantiles of Chronos-2 under three calibrations.

Policy	Thermal (1.20)	Storage (1.02)	Compromise (1.18)
Chronos-2 native median	262	51	277
Chronos-2 native 0.4 quantile	173	61	188
Chronos-2 native 0.3 quantile	91	83	138
Chronos-2 native 0.25 quantile	71	100	164
Chronos-2 native 0.2 quantile	111	132	223
Chronos-2 native 0.25 quantile, comonotonic	213	321	375
Chronos-2 native quantile chosen from past years	265	51	279
Chronos-2 debiased, −10%	71	92	150
PAR(p) recalibrated, debiased, −10%	124	72	122
PAR(p) recalibrated, empirical 0.25 quantile	71	112	178
Chronos-2, empirical 0.25 quantile	87	164	213
PAR(p) recalibrated, debiased, no hedging	308	54	292

Regret in R$ billion of December 2024 relative to the perfect-inflow policy of each calibration, observed history 2005 to 2024.

Expected performance over resampled histories

With forecasts re-run on each of the 100 resampled histories, hedging was the robust result when the system was scarce and the choice of forecaster a smaller, second-order one (Table 9, Figure 8). Under the thermal calibration a 10% haircut lowered the expected regret of the debiased PAR(p) from R$ 239 billion to 110, a difference of R$ 129 billion (95% interval 104 to 155), and was cheaper in 88% of histories; for the debiased Chronos-2 it lowered regret by R$ 120 billion (96 to 145), cheaper in 85%. The compromise calibration gave almost the same picture (R$ 120 billion, 88%). Under the storage calibration the haircut still lowered the expected regret of PAR(p), by R$ 29 billion (16 to 43), but was cheaper in only 52% of histories, and for Chronos-2 the effect was indistinguishable from zero (-5, -17 to 8). Insurance pays on average where droughts can make the system short, and is close to a coin toss where they rarely do.

At equal haircut the debiased Chronos-2 was cheaper than the debiased PAR(p) in expectation under the three calibrations, by R$ 16 billion (8 to 24) under the thermal calibration, 15 billion (7 to 22) under the compromise and 10 billion (1 to 19) under the storage calibration, and it was cheaper in 66%, 66% and 37% of histories. Without hedging the differences were larger, R$ 25 to 35 billion, and the uncorrected Chronos-2, whose forecasts are lower, was cheaper than PAR(p) by R$ 114 to 120 billion in the scarce calibrations. The native 0.25 quantile of Chronos-2 was cheaper than PAR(p) with a 10% haircut by R$ 23 billion (9 to 39) under the thermal calibration and 14 billion (1 to 29) under the compromise, and was close to Chronos-2 with the same haircut in the scarce calibrations. The difference between forecasters at equal hedging was therefore real in expectation but about a tenth of the effect of hedging in the scarce system, and it changed sign from one history to the next.

Whether forecasts are re-run or transported mattered for these comparisons. Transported forecasts gave lower regrets than re-run ones for every policy, by R$ 21 to 88 billion, because they keep the accuracy each model achieved on the real sequence, whereas a model run on a synthetic history faces block boundaries and sequences it has never seen. For the effect of hedging the two

methods agreed in sign in 80% to 84% of histories, and transported forecasts understated it (R$ 103 against 129 billion under the thermal calibration). For the comparison between forecasters at equal haircut they disagreed: transported forecasts made PAR(p) cheaper by R$ 9 to 15 billion in all three calibrations, the opposite of the re-run result, and the two methods agreed in sign in only 40% to 77% of histories, with correlations of 0.10 to 0.52 (Table 10). Transported forecasts are therefore a usable sensitivity analysis for hedging but not a basis for ranking forecasters, which requires running the models again.

Table 9. Decision experiment over 100 resampled histories with re-run forecasts, three calibrations.

Quantity	Thermal (1.20)	Storage (1.02)	Compromise (1.18)
Regret, PAR(p) rec., debiased	239 (208 to 270)	132 (109 to 156)	232 (205 to 260)
Regret, PAR(p) rec., debiased, −10%	110 (93 to 127)	103 (90 to 116)	111 (95 to 128)
Regret, Chronos-2, debiased	214 (183 to 247)	97 (82 to 116)	198 (172 to 228)
Regret, Chronos-2, debiased, −10%	94 (76 to 112)	93 (82 to 104)	97 (80 to 114)
Regret, Chronos-2 native 0.25 quantile	87 (68 to 106)	105 (95 to 114)	97 (78 to 115)
Effect of a 10% haircut on PAR(p)	−129 (−155 to −104); 88%	−29 (−43 to −16); 52%	−120 (−145 to −97); 88%
Effect of a 10% haircut on Chronos-2	−120 (−145 to −96); 85%	−5 (−17 to 8); 37%	−102 (−126 to −79); 80%
Chronos-2 deb. − PAR(p) deb., no haircut	−25 (−38 to −13); 70%	−35 (−47 to −24); 58%	−33 (−46 to −21); 70%
Chronos-2 deb. − PAR(p) deb., both −10%	−16 (−24 to −8); 66%	−10 (−19 to 0); 37%	−15 (−22 to −7); 66%
Native 0.25 quantile − PAR(p) deb., −10%	−23 (−39 to −9); 56%	+2 (−11 to 14); 32%	−14 (−29 to −1); 54%
Native 0.25 quantile − Chronos-2 deb., −10%	−7 (−25 to 10); 38%	+12 (4 to 18); 19%	0 (−16 to 16); 38%

Mean over histories in R$ billion of December 2024 with 95% bootstrap intervals over histories; regret is relative to the perfect-inflow policy of each history. For differences a − b, the percentage is the share of histories in which a was cheaper.

Table 10. Comparisons over resampled histories with re-run and with transported forecasts.

Comparison	Calibration	Re-run	Transported	Same sign	Corr.
Effect of a 10% haircut on PAR(p)	Thermal	−129 (−155 to −104)	−103 (−126 to −80)	0.80	0.65
Effect of a 10% haircut on PAR(p)	Storage	−29 (−43 to −16)	−15 (−29 to −2)	0.84	0.73

Comparison	Calibration	Re-run	Transported	Same sign	Corr.
Effect of a 10% haircut on PAR(p)	Compromise	−120 (−145 to −97)	−80 (−101 to −60)	0.82	0.61
Chronos-2 − PAR(p), no haircut	Thermal	−25 (−38 to −13)	−27 (−36 to −17)	0.63	0.07
Chronos-2 − PAR(p), no haircut	Storage	−35 (−47 to −24)	−2 (−8 to 3)	0.74	0.51
Chronos-2 − PAR(p), no haircut	Compromise	−33 (−46 to −21)	−21 (−31 to −13)	0.65	0.09
Chronos-2 − PAR(p), both −10%	Thermal	−16 (−24 to −8)	+9 (4 to 14)	0.40	0.10
Chronos-2 − PAR(p), both −10%	Storage	−10 (−19 to 0)	+15 (11 to 19)	0.77	0.52
Chronos-2 − PAR(p), both −10%	Compromise	−15 (−22 to −7)	+11 (8 to 16)	0.53	0.32

Mean differences a − b in R$ billion over 100 histories with 95% bootstrap intervals. Same sign: share of histories in which the two methods give a difference of the same sign; Corr.: correlation across histories.

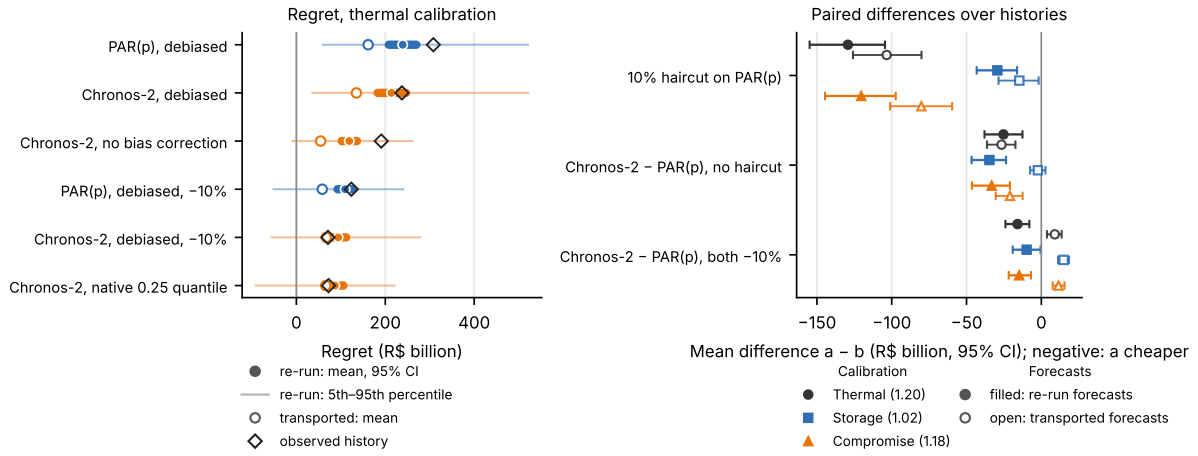


Figure 8. Decision experiment over 100 resampled hydrological histories. Left: regret under the thermal calibration; filled circles with thick bars are the mean and 95% interval with re-run forecasts, thin lines span the 5th to 95th percentiles across histories, open circles are the mean with transported forecasts and diamonds the observed history. Right: mean paired differences a − b with 95% intervals under the three calibrations, distinguished by marker and colour; filled markers use re-run forecasts and open markers transported forecasts, and negative values mean that a is cheaper.

Discussion

The three evaluations tell a consistent story when read in order. In accuracy, the foundation models are competitive and, at long leads, the best. Weighted by the price of energy, they

have less exposure, but one month ahead that advantage is their bias: they under-forecast, and under-forecasting is what a hydrothermal operator wants when water is scarce. In the decision experiment, every comparison, between forecasters or between hedging levels, turns on a few months of scarcity. Over the observed history the drought of 2021 carried more than the whole difference in each case, and without it the sign of the comparison reversed. Over histories resampled from the same years, in which droughts arrive at different times and with different frequency, and with the models run again on each, the effect of hedging became clear when the system was scarce and close to neutral when it was slack. Hedged alike, the foundation model was cheaper in expectation by an amount about a tenth of the hedging effect, and cheaper in only one to two thirds of the histories. None of this establishes that Chronos-2 is operationally superior to a recalibrated autoregressive model: over the record that actually happened no difference between them was statistically clear, and the small expected advantage depends on how the synthetic histories are built.

This has a simple reading. Under stored water and a costly thermal margin, a forecast or a hedge earns its value in the rare months in which reservoirs run low and loses a little in all the others. Hedging behaves like insurance: the premium, in fuel and emissions, is paid every year and is well determined, while the payout comes in a drought whose timing cannot be learned from a short record, as the nested rule showed by adopting a haircut only after 2021. Over the resampled histories the insurance paid in 88% of them when the simulated system was scarce and in about half when it was slack. A calibrated predictive distribution, such as the native quantiles of Chronos-2 aggregated with the dependence between basins, provides the same insurance without an arbitrary haircut, but it does not tell the operator which quantile to use; that choice is a statement about risk aversion and scarcity, not about forecast skill. A pessimistic forecaster behaves the same way, which is why Chronos-2 looked valuable in the price-weighted analysis and in the closed loop until its bias was removed or the drought year excluded.

For forecast evaluation in power systems, the lessons are practical. Comparisons by NSE, or by error weighted by an observed price, reward the sign of the error as much as its size, so forecasters should be compared after removing their bias and with the same hedging rule. Closed-loop comparisons over a single historical record should report how much of each difference comes

from individual years and should test equivalence rather than read a wide interval as equality. And because one drought can decide the result, expected performance should be estimated over many plausible hydrological histories, of the kind used in stochastic planning [1, 2], and not only over the one that happened. Those comparisons need the models to be run again on each history: transporting the errors a model made on the real sequence was adequate for the effect of hedging but reversed the comparison between forecasters. Histories drawn from stochastic inflow models [24], with droughts beyond the observed ones, would test the result further. Refitting the operational model every year and removing its bias costs nothing and lowered price-weighted exposure and simulated regret in every comparison we made; the further step to a foundation model was not distinguishable from noise over the observed record once both were treated alike, and was small in expectation.

These results come from one system, one twenty-year record and one simplified simulator, and they support a hypothesis rather than a general law: that in storage-rich hydrothermal systems the value of forecasts and of hedging concentrates in rare droughts. How far this travels depends on how much storage a system has relative to its inflow and on how expensive a shortfall is. The SIN combines multi-year storage, a large thermal fleet that is costly at the margin and long dry periods, which is the setting in which value concentrates in rare droughts and a bias towards less water pays when one arrives. Systems with large storage and a liquid market, such as the Nordic one, share the first two features and might show the same concentration, with market prices rather than a centrally computed marginal cost carrying the signal. In snow-dominated systems such as the Columbia River, seasonal volume forecasts are skilful and carry substantial economic value [10, 11]; with limited carry-over storage, the information in the forecast, rather than its bias, is more likely to matter. Where most capacity is run-of-river and storage is small, decisions are short-term and forecast accuracy at days to weeks should weigh more than monthly hedging. Our calibrations point in the same direction within one system: the slacker the simulated system, the less hedging was worth.

The simulator limits the strength of every closed-loop result. One equivalent reservoir without transmission limits, inflexible thermal generation or plant-level constraints moves water more freely than the real system, which is why demand had to be inflated by 20% to reproduce

observed thermal generation; with that calibration the simulated system was drier than the real one from 2016 to 2021 and the 2021 drought was more severe, and the calibration on storage, which matched the reservoirs, left almost no scarcity at all. No calibration met more than two of the five adequacy criteria, and the economic results should be read as conditional on the scarcity each calibration implies; a representation by subsystems with transmission limits would be the natural next step. A deterministic rolling-horizon rule with a guide curve, rather than the stochastic and risk-averse programming the operator uses [2, 16], is myopic beyond twelve months, which is why the hindsight optimum was R$ 178 billion cheaper than the perfect-inflow policy and why a hedge can partly substitute for missing risk aversion. These simplifications inflate the absolute regret and the absolute value of hedging, and they concentrate it in the driest years; the costs in Table 4 indicate scale and should not be read as what each forecast would have cost Brazilian consumers. The relative comparison between forecasters at equal hedging was never statistically clear over the observed record and was small over the resampled histories, and the value of hedging changed with scarcity. The resampled histories reuse the observed droughts, so they do not test the forecasters on droughts more severe than 2021. Nearly half of the plant-month thermal capacities were imputed; the imputed share was not concentrated in the driest years, but removing or rescaling those plants moved results as much as any change of model. Emissions use typical factors per fuel and a partial fuel match, although their difference between policies was robust to the factors. The predictive distributions were used through a single quantile in a deterministic rule; feeding them to the scenario generation of a stochastic planning model would use them more fully, and a single system was studied.

Conclusions

Over twenty years of monthly inflow forecasts for the Brazilian interconnected system, time-series foundation models were competitive in accuracy and had less price-weighted exposure than a recalibrated and bias-corrected periodic autoregressive model, but one month ahead that advantage came from a pessimistic bias. In the closed-loop decision experiment, with information available at each decision, neither the choice of forecaster nor the level of hedging could be distinguished over the observed history, because more than the whole of each difference

came from the 2021 drought; a hedge chosen from past years alone was adopted only after it, and the emissions cost of hedging was certain while its benefit was not. With both models re-run on 100 histories resampled from the same years, a 10% haircut paid in 88% of them when the simulated system was scarce and in about half when it was slack, while the foundation model, hedged alike, was cheaper in expectation by about a tenth of that effect; forecasts transported from the observed record were adequate for the effect of hedging but not for ranking forecasters. The native predictive distribution of Chronos-2 was well calibrated one month ahead and hedged as well as the best fixed haircut, but the right quantile depended on scarcity. These conclusions are conditional on a simplified decision experiment that met only part of the adequacy criteria, and its simulated costs indicate scale rather than real costs. Forecasting models for hydrothermal systems should be compared after bias correction, under the same hedging rule and against a recalibrated reference, with cost, deficit and emissions reported together with the dependence of any ranking on individual droughts made explicit, and over many hydrological histories, with the models run again on each, rather than the observed one alone.

Supplementary information

Additional file 1 (xlsx). Table S1: thermal capacity imputed for unmatched plants by year, and share of thermal generation supplied by them. Table S2: imputation by fuel. Table S3: adequacy criteria of the simulator over the grid of demand factors. Table S4: software and model versions, random seeds and parameters of every experiment.

List of abbreviations

CMO: marginal operation cost (custo marginal de operação); CVU: variable unit cost of a thermal plant; ENA: natural energy inflow; GWmed: average gigawatts over a period; IPCA: Brazilian consumer price index; LP: linear programme; MAE: mean absolute error; NSE: Nash-Sutcliffe efficiency; ONS: Brazilian National Grid Operator; PAR(p): periodic autoregressive model of order p; SIN: Brazilian National Interconnected System.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analysed during the current study are publicly available: ONS open data (https://dados.ons.org.br, CC-BY licence) for daily natural inflow, ENA, stored energy, the energy balance, the registers of reservoirs and installed capacity, CVU and CMO, and the IPCA series 433 of the Central Bank of Brazil. The inflow forecasts generated and the code that downloads the data, converts the forecasts, runs the dispatch simulations and the statistics, and produces every table and figure are available in the hydrobench repository, https://github.com/leofome8/hydrobench, with one notebook per stage of the analysis; model versions, random seeds, copula settings and the other parameters of every experiment are listed in Additional file 1 (Table S4).

Competing interests

The authors declare that they have no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

LF conceived the study, designed the methodology, wrote the software, carried out the formal analysis and data curation, prepared the figures and wrote the first draft. HDMJ and AGGL

validated the methods and results and reviewed and edited the manuscript. MACP supervised the work and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Pereira MVF, Pinto LMVG. Multi-stage stochastic optimization applied to energy planning. Math Program. 1991;52:359–75.
2. Maceira MEP, Penna DDJ, Diniz AL, Pinto RJ, Melo ACG, Vasconcellos CV, et al. Twenty years of application of stochastic dual dynamic programming in official and agent studies in Brazil: main features and improvements on the NEWAVE model. In: 2018 Power Systems Computation Conference (PSCC). Dublin: IEEE; 2018. p. 1–7.
3. Salas JD, Delleur JW, Yevjevich V, Lane WL. Applied modeling of hydrologic time series. Littleton: Water Resources Publications; 1980.
4. Maceira MEP, Damázio JM. Use of the PAR(p) model in the stochastic dual dynamic programming optimization scheme used in the operation planning of the Brazilian hydropower system. Probab Eng Inf Sci. 2006;20:143–56.
5. Kratzert F, Klotz D, Herrnegger M, Sampson AK, Hochreiter S, Nearing GS. Toward improved predictions in ungauged basins: exploiting the power of machine learning. Water Resour Res. 2019;55:11344–54.
6. Ansari AF, Stella L, Turkmen C, Zhang X, Mercado P, Shen H, et al. Chronos: learning the language of time series. Trans Mach Learn Res. 2024.
7. Ansari AF, Shchur O, Küken J, Auer A, Han B, Mercado P, et al. Chronos-2: from univariate to universal forecasting. arXiv:2510.15821. 2025.
8. Nash JE, Sutcliffe JV. River flow forecasting through conceptual models part I: a discussion of principles. J Hydrol. 1970;10:282–90.
9. Murphy AH. What is a good forecast? An essay on the nature of goodness in weather

forecasting. Weather Forecast. 1993;8:281–93.

10. Hamlet AF, Huppert D, Lettenmaier DP. Economic value of long-lead streamflow forecasts for Columbia River hydropower. J Water Resour Plan Manag. 2002;128:91–101.
11. Anghileri D, Voisin N, Castelletti A, Pianosi F, Nijssen B, Lettenmaier DP. Value of long-term streamflow forecasts to reservoir operations for water supply in snow-dominated river catchments. Water Resour Res. 2016;52:4209–25.
12. Matte S, Boucher MA, Boucher V, Fortier Filion TC. Moving beyond the cost–loss ratio: economic assessment of streamflow forecasts for a risk-averse decision maker. Hydrol Earth Syst Sci. 2017;21:2967–86.
13. Cassagnole M, Ramos MH, Zalachori I, Thirel G, Garçon R, Gailhard J, et al. Impact of the quality of hydrological forecasts on the management and revenue of hydroelectric reservoirs: a conceptual approach. Hydrol Earth Syst Sci. 2021;25:1033–52.
14. Hong T, Pinson P, Fan S, Zareipour H, Troccoli A, Hyndman RJ. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. Int J Forecast. 2016;32:896–913.
15. Hong T, Pinson P, Wang Y, Weron R, Yang D, Zareipour H. Energy forecasting: a review and outlook. IEEE Open Access J Power Energy. 2020;7:376–88.
16. Shapiro A, Tekaya W, da Costa JP, Soares MP. Risk neutral and risk averse stochastic dual dynamic programming method. Eur J Oper Res. 2013;224:375–91.
17. Rockafellar RT, Uryasev S. Optimization of conditional value-at-risk. J Risk. 2000;2(3):21–41.
18. Operador Nacional do Sistema Elétrico (ONS). Portal de dados abertos. https://dados.ons.org.br. Accessed 22 Oct 2026.
19. Nie Y, Nguyen NH, Sinthong P, Kalagnanam J. A time series is worth 64 words: long-term forecasting with transformers. In: International Conference on Learning Representations (ICLR); 2023.
20. Vogel RM, Shallcross AL. The moving blocks bootstrap versus parametric time series models. Water Resour Res. 1996;32:1875–82.
21. Künsch HR. The jackknife and the bootstrap for general stationary observations. Ann Stat.

1989;17:1217–41.

22. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6:65–70.

23. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm. 1987;15:657–80.

24. Treistman F, Maceira MEP, Penna DDJ, Damázio JM, Rotunno Filho OC. Synthetic scenario generation of monthly streamflows conditioned to the El Niño–Southern Oscillation: application to operation planning of hydrothermal systems. Stoch Environ Res Risk Assess. 2020;34:331–53.