

Epidemic Forecasting: Lessons Learned from the SARS-CoV-2 Pandemic to Balance Accuracy, Feasibility, and Impact

Thomas Ferté^{1,2}, Vincent Thevenet^{2,3}, Xavier Hinaut^{4,5,6}, Pierrick Legrand^{4,7,8}, Dan Dutartre⁴, Romain Griffier^{2,9}, Viannet Jouhet^{2,9}, Boris P. Hejblum¹, and Rodolphe Thiébaud^{1,2,*}

¹Univ. Bordeaux, INSERM, Inria, BPH, U1219, équipe SISTM, F-33000 Bordeaux, France

²CHU de Bordeaux, Pôle de santé publique, Service d'information médicale, F-33000 Bordeaux, France

³Univ. Bordeaux, INSERM, BPH, U1219, équipe EPICENE, F-33000 Bordeaux, France

⁴INRIA Centre of Bordeaux University, France

⁵LaBRI, Bordeaux University, Bordeaux INP, CNRS UMR 5800, France

⁶IMN, Bordeaux University, CNRS UMR 5293, Bordeaux, France

⁷Bordeaux INP, ENSC, F-33400 Talence, France

⁸Univ. Bordeaux, CNRS, Bordeaux INP, IMS, UMR 5218, F-33400 Talence, France

⁹Univ. Bordeaux, INSERM, BPH, U1219, équipe AHead, F-33000 Bordeaux, France

*Corresponding author: rodolphe.thiebaut@u-bordeaux.fr

Abstract

The COVID-19 pandemic highlighted the importance of reliable, real-time hospital forecasting. At Bordeaux University Hospital, we developed models to predict SARS-CoV-2-related hospitalizations 14 days in advance using integrated data sources. We identified six key lessons to guide future epidemic response: (1) Multimodal data improves accuracy; (2) Simple baseline models are essential for benchmarking and building trust; (3) Model and metric choices must align with decision goals which often means prioritizing absolute over relative metrics and beginning with simple models; (4) Prediction intervals should be pro-

vided to communicate the uncertainty associated with forecasts; (5) Real-world constraints such as computational cost, maintainability, and required expertise should guide model selection; (6) Forecasts must be contextualized and communicated carefully to policymakers. We advocate for a systems-level forecasting approach that balances accuracy, feasibility, and impact.

1 Introduction

The COVID-19 pandemic, caused by SARS-CoV-2, emerged in late 2019 and quickly became a global health crisis (1, 2). Countries implemented enhanced surveillance, non-pharmaceutical interventions (e.g., lockdowns, curfews, school closures), and mass vaccination (3–6). Hospitals faced unprecedented pressure, prompting measures such as postponing elective procedures (7).

Anticipating epidemic trajectories was a major challenge, affecting timely and efficient care (8). Short-term forecasting models emerged but were limited by methodological heterogeneity, lack of standardization, and variable operational utility. Approaches ranged from simple extrapolation to advanced machine learning, often without clear guidance on comparative advantages (9, 10).

At Bordeaux University Hospital (BUH), our multidisciplinary team developed a real-time forecasting model using penalized linear regression on hospital, public health, and weather data (11), later exploring machine learning and neural networks to improve predictions (12–14). Beyond algorithmic performance, we faced practical challenges including forecast reliability, evaluation metrics, and communication with decision-makers.

This article provides an experience-based perspective on developing and deploying real-time epidemic forecasts, highlighting practical challenges, key decisions, and lessons for more actionable forecasting. Section 2 introduces the setting and models, followed by insights on data quality (Section 3), baseline models (Section 4), modeling approaches (Section 5), prediction uncertainty (Section 6), operational constraints (Section 7) and communication with policymakers (Section 8).

2 Methods

We illustrate our discussion using forecasting efforts during the COVID-19 epidemic at BUH, aiming to predict daily SARS-CoV-2 hospitalizations at a 14-day horizon. We chose a 14-day horizon to give the hospital enough time to respond while keeping forecasts practical (Section 4.2). Models were trained and evaluated every two days using integrated data from BUH’s Electronic Health Records, public COVID-19 statistics, weather, and vaccination coverage (15–17), spanning May 16, 2020, to January 17, 2022. The dataset included 409 predictors over 586 days. We compared three models:

- **Elastic net (ENet):** A simple, interpretable statistical model, easy to implement.

- **XGBoost:** A widely used tree-based machine learning algorithm for time series forecasting (9, 18).
- **Reservoir Computing (RC):** A neural network approach previously shown to perform best in our benchmarks (12). RC uses a randomly initialized recurrent “reservoir” layer, training only the output layer to capture complex temporal dependencies efficiently (19–22). Hyperparameters were optimized using a Genetic Algorithm (12). RC has been applied in diverse domains, including epidemic modeling, language processing, and financial forecasting (23–28).

Heavier models such as Transformers showed poor performance in this context (12) and were excluded. Additional methodological details, including pre-processing, model specifications, hyperparameter ranges, and feature selection, are provided in Supplementary Materials.

3 Key Point 1: Choosing the Right Data Sources

3.1 Multi-modal Data Integration

Comito et al. (9) reviewed SARS-CoV-2 forecasting features, including daily cases, deaths, recoveries, vaccination rates, and phone calls. They found no consensus on which features should be prioritized, likely due to local epidemic variability and prediction model differences. Similarly, Cramer et al. (29) observed no consistent benefit from using all available data sources in their benchmark. Nevertheless, most effective models integrate diverse data types.

Weather. Environmental factors like temperature and humidity influence viral spread (30, 31). Roumagnac et al. (31) proposed the IPTCC index, which scores transmission potential from 0 to 100 to capture this phenomenon. In our work, incorporating weather into 14-day forecasts improved accuracy: Median Absolute Error (MdAE) for Enet regression dropped from 17.21 to 11.63 (11), and RC Mean Absolute Error (MAE) decreased from 15.67 to 15.27 (12).

Electronic Health Records (EHRs). Hospital data warehouses enable epidemic forecasting from real-time EHR streams, predicting admissions within hours or weekly visit volumes (11, 32, 33). Aggregated EHR data in our models reduced 14-day forecast MdAE from 11.63 to 9.48 hospitalizations (11).

Mobility. Mobility data from platforms such as Facebook, Google, and Baidu were used during the pandemic to capture population movement patterns. Several studies demonstrated their utility in improving forecast accuracy (34–36).

Digital Surveillance. Digital epidemiology tools, such as Google Trends and social media platforms, have been explored for epidemiological surveillance (37). However, findings in the context of COVID-19 have been mixed, due to the strong influence of media coverage and public announcements on search behavior, as well as variable lag times between symptom searches and case incidence (38, 39).

Wastewater-Based Surveillance (WBS). WBS provides a non-invasive method to monitor both symptomatic and asymptomatic SARS-CoV-2 infections by quantifying viral load in sewage. Unlike traditional testing, it captures a broader, more representative population signal. Studies have highlighted its early-warning potential, although timely sample processing is critical for real-time forecasting (40–43).

Data Validity and Standardization Across Data Sources. Beyond obtaining accurate data, a major challenge is the lack of standardized definitions across sources. For example, reported COVID-19 hospitalizations may vary widely depending on institutional criteria, particularly regarding when a patient is considered no longer infectious, underscoring the need for harmonized reporting to enhance forecasting accuracy.

3.2 Data Timeliness and Updating Frequency

Latency between events and database registration affects forecasting. Hospital data warehouses often provide faster updates than public sources. At BUH, clinical data were available within one day, compared with four-day delays in traditional epidemiologic surveillance. This faster access improved 14-day hospitalization forecast MdAE by 2.60 points (11).

4 Key Point 2: Setting a Baseline Model

4.1 A Not-So-Trivial Baseline

A simple yet informative baseline model can be derived from available data. In our context, this baseline forecasts 14-day hospitalizations by assuming the value remains unchanged, essentially following the last observation carried forward (LOCF) method.

Eventhough outperforming such a naive baseline may seem trivial, empirical evidence suggests otherwise. Cramer et al. (29) evaluated 28 models predicting incident COVID-19 deaths. Only 18 out of 28 models achieved a Mean Absolute Error equal to or lower than this simple baseline. Similarly, Bracher et al. (44) found that 6 out of 8 models failed to outperform the baseline when forecasting two-week cumulative cases in Germany.

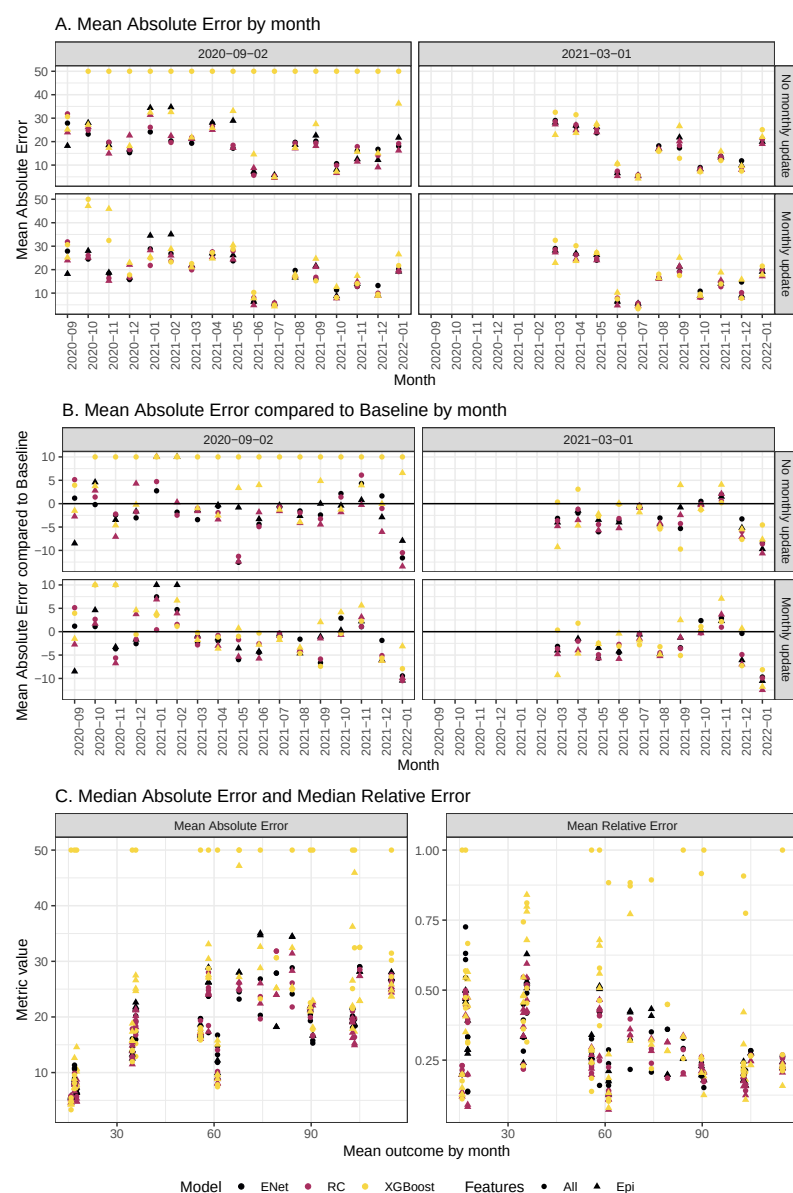


Figure 1: Panels A and B show MAE and MAEB by month. The first column shows models trained since 2020-09-02, and the second column shows models trained since 2021-03-01. The first row shows models without monthly updates, and the second row shows models with monthly updates. Panel C shows MAE and MAEB as a function of the mean outcome by month. MAE values above 50 were capped at 50, and MRE values above 1 were set to 1 for visualization purposes.

4.2 Why is this Baseline Model Useful?

Detecting ineffective models. Models that fail to outperform the baseline add no predictive value and are unlikely to support decision-making. In our setting, Transformer models exhibited this behavior (12).

Defining forecast horizons. Neither XGBoost nor RC models outperformed the baseline for 21-day forecasts, suggesting that this horizon may be out of reach. Focus should instead be placed on shorter horizons (e.g., 7- or 14-day) (12). This aligns with prior findings that even 14-day forecasts are challenging, with typical delays of about six days between observed slope shifts and corresponding prediction changes (11).

Assessing data sufficiency. Comparing model performance to this baseline over time provides key insights. Panel A of Figure 1 shows MAE typically around 20 hospitalizations, an arguably acceptable error margin. However, Panel B indicates that the Mean Absolute Error relative to the baseline (MAEB, i.e. the mean absolute difference between model error and baseline error) was often above zero before March 2021, meaning models frequently performed worse than the baseline during the early epidemic phase. This suggests that early in the pandemic, no model offered predictive value beyond the last observed data point.

In summary, although straightforward to implement, this baseline model provides a key benchmark. It sets a minimum standard for forecasting performance that models must exceed, guiding model selection, informing forecast reliability assessments, and setting realistic expectations for predictive accuracy.

5 Key Point 3: Finding the best model

5.1 The right metric for the right task

Comito et al. (9) found 14 different metrics used across forecasting studies, with most reporting multiple metrics and few designating a primary one. Exceptions include Zheng et al. (45) and Gerlee et al. (46), who emphasized Mean Absolute Percentage Error (MAPE), though the rationale for prioritizing it over absolute metrics like MAE is unclear.

Choice of metric matters because different metrics emphasize different performance aspects and affect model selection. Absolute metrics (e.g., MAE, MSE) measure deviations in raw hospitalization counts, while relative metrics (e.g., MAPE, Laplacian Error) scale errors by observed values, effectively down-weighting high-incidence deviations.

Our goal was to forecast hospital occupancy during epidemic peaks, when underestimating hospitalizations risks capacity strain and delayed care. Metrics sensitive to peak errors are therefore preferable. Figure 1 shows that MAE rises with hospitalization counts, capturing large deviations during peaks, whereas

Median Relative Error (MRE) down-weights them. For anticipating hospital surges, MAE is more appropriate.

In summary, aligning metrics with forecasting objectives ensures model evaluation supports real-world decision-making.

5.2 Data preparation

Preprocessing is often beneficial before feeding data into a model. In the context of COVID-19 forecasting, three key aspects are smoothing, computing derivatives, and transforming outcomes.

Smoothing. Because COVID-19 data are highly variable over time, smoothing can facilitate model learning. In Ferté et al. (11), local polynomial smoothing reduced the MdAE by 2.47 hospitalizations compared to no smoothing. Paireau et al. (34) also showed that it outperformed a simple 7-day moving average.

Derivatives. Computing derivatives can also improve performance. In linear regression, derivatives reduced MdAE by 3.24 hospitalizations (11). Even with more complex models such as RC, explicitly including derivatives reduced MAE by 3.47 hospitalizations (12), suggesting that even complex models do not always extract this information spontaneously.

Outcome transformations. Learning directly from raw outcomes is not always optimal. Alternatives include learning growth rates (34) or applying log-transformations (47), both based on the assumption of local exponential growth. In our work, we proposed learning the residuals of the LOCF baseline relative to hospitalizations rather than raw hospitalization counts, thereby simplifying the task (14). In other words, we modeled the change in hospitalizations rather than their absolute values.

5.3 Statistical learning vs Machine Learning

Several articles have explored the distinctions and overlap between statistical learning and machine learning (48–50), including in the context of biopreparedness (51). Statistical learning is typically valued for its interpretability, simplicity, and speed, though it often relies on linear assumptions and struggles with high-dimensional data. In contrast, machine learning handles complex, non-linear patterns and high-dimensional inputs better, but is generally less interpretable and more computationally intensive.

This contrast is not absolute. Feature engineering can address some limitations of statistical models, whose interpretability also declines with more and correlated features. Meanwhile, machine learning has become more accessible thanks to modern software, and its computational demands are increasingly manageable.

Table 1: Performance by algorithm for models initiated on 2020-09-02 and evaluated from 2020-09-02 to 2021-03-15.

Model	Features	Update	MAE ($\pm$ SD)	MAEB ($\pm$ SD)
Baseline	-	-	21.03($\pm$ 12.45)	0.00($\pm$ 0.00)
ENet	all	No	20.47($\pm$ 13.86)	-0.56($\pm$ 14.04)
	all	Yes	22.12($\pm$ 15.73)	1.09($\pm$ 14.23)
	epi	No	24.36($\pm$ 20.26)	3.33($\pm$ 17.76)
	epi	Yes	24.65($\pm$ 20.54)	3.61($\pm$ 18.33)
RC	all	No	21.54($\pm$ 18.39)	0.51($\pm$ 19.52)
	all	Yes	20.82($\pm$ 18.42)	-0.21($\pm$ 17.7)
	epi	No	22.47($\pm$ 18.41)	1.44($\pm$ 18.5)
	epi	Yes	22.69($\pm$ 16.69)	1.66($\pm$ 17.0)
XGBoost	all	No	161.48($\pm$ 253.34)	140.44($\pm$ 255.67)
	all	Yes	28.72($\pm$ 24.91)	7.68($\pm$ 23.19)
	epi	No	24.95($\pm$ 19.65)	3.91($\pm$ 19.25)
	epi	Yes	31.43($\pm$ 37.04)	10.39($\pm$ 36.03)

The two paradigms also differ in approach: epidemiology-driven modeling emphasizes expert feature selection, while machine learning favors a more data-driven, holistic strategy. As Beam et al. (50) suggests, these approaches form a continuum, where more data often calls for more automated methods.

To examine these differences, we selected three representative methods: Enet for statistical modeling, XGBoost for tree-based machine learning, and RC for neural network-based learning. We then explore the following hypotheses: (1) machine learning requires more data than statistical models; (2) machine learning can outperform statistical models due to fewer assumptions and (3) expert-driven feature selection outperforms purely data-driven selection.

5.3.1 Quantity of data needed

We first examined a setting with limited information (2020-09-02 to 2021-03-15, Table 1). We hypothesized that simpler models would outperform complex ones, which risk overfitting. Indeed, linear regression with expert-selected features achieved the lowest MAE. However, the MAEB metric showed only a marginal improvement over the LOCF baseline (0.56 hospitalizations), a clinically negligible difference (Figure 2). Overall, when data are limited, neither statistical nor machine learning models provide accurate forecasts.

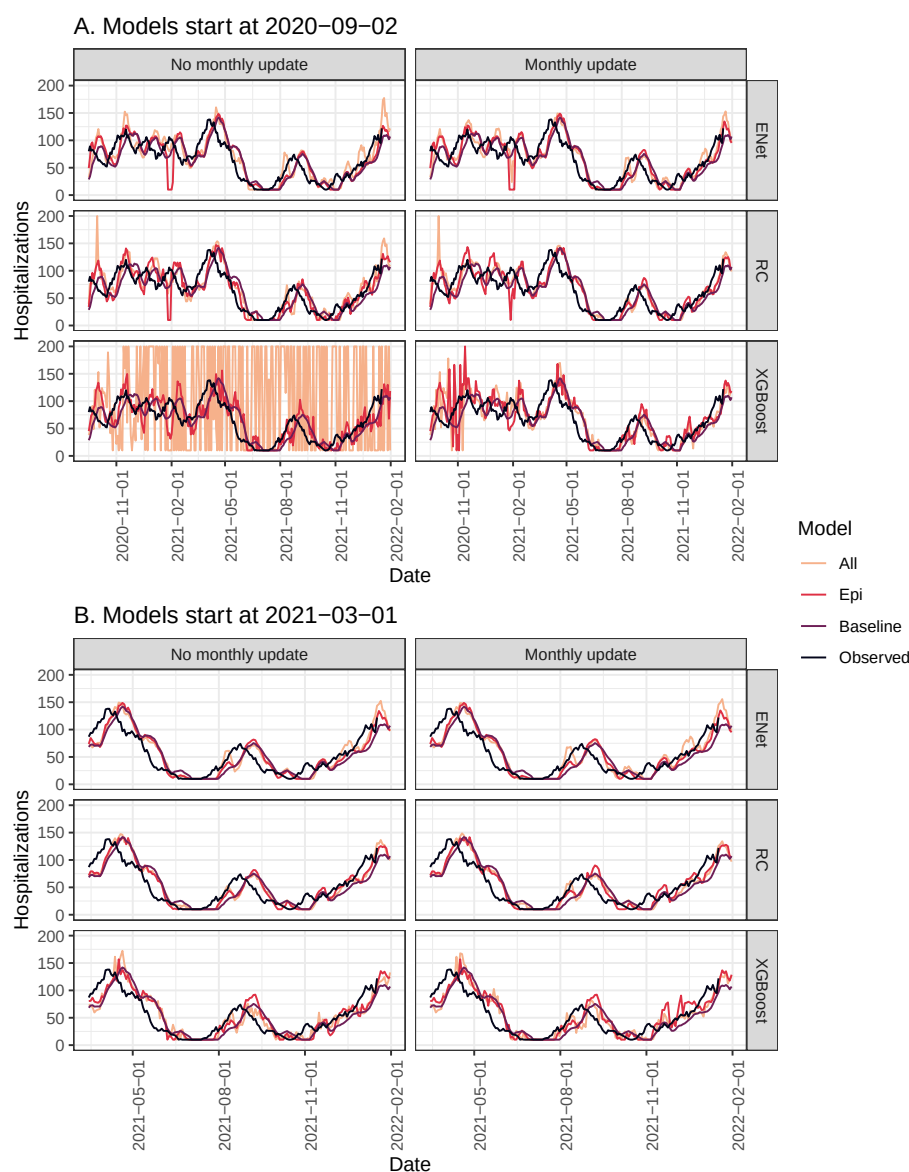


Figure 2: Forecasts generated by different models initialized at various start dates, with or without monthly hyperparameter updates, using either all available features (All) or a feature set selected based on a priori knowledge (Epi).

Table 2: Performance by algorithm for models initiated on 2021-03-01 and evaluated from 2021-03-15 to 2022-01-17. Each algorithm was repeated three times; the table shows the median MAE and MAEB along with their minimum and maximum values.

Model	Features	Update	MAE (min ; max)	MAEB (min ; max)
Baseline	-	-	18.59 (18.59 ; 18.59)	0.00 (0.00 ; 0.00)
ENet	all	No	15.83 (15.82 ; 15.87)	-2.76 (-2.77 ; -2.72)
	all	Yes	15.98 (15.9 ; 16.29)	-2.6 (-2.68 ; -2.29)
	epi	No	15.93 (15.86 ; 16)	-2.65 (-2.73 ; -2.58)
	epi	Yes	15.88 (15.87 ; 15.89)	-2.7 (-2.72 ; -2.7)
RC	all	No	15.4 (15.24 ; 15.6)	-3.19 (-3.35 ; -2.99)
	all	Yes	15.63 (15.59 ; 15.65)	-2.96 (-3 ; -2.93)
	epi	No	15.16 (15.05 ; 17.45)	-3.43 (-3.53 ; -1.14)
	epi	Yes	15.1 (14.96 ; 15.46)	-3.48 (-3.62 ; -3.12)
XGBoost	all	No	16.08 (15.98 ; 19.95)	-2.51 (-2.61 ; 1.36)
	all	Yes	16.29 (16.26 ; 16.78)	-2.29 (-2.33 ; -1.8)
	epi	No	16.44 (16.42 ; 16.97)	-2.15 (-2.17 ; -1.62)
	epi	Yes	16.99 (16.61 ; 17.32)	-1.59 (-1.97 ; -1.27)

5.3.2 Models comparison

We then assessed whether machine learning outperforms statistical learning with more data (Table 2). All models exceeded the baseline by about 2.5 hospitalizations. XGBoost performed worse than both RC and ENet, while RC slightly outperformed ENet by less than one hospitalization.

These results highlight that statistical learning remains competitive with machine learning approaches. Also, as expected, all machine learning approaches are not behaving the same and careful choice should be done among these approaches. In our setting, the last neural network layer of RC is a linear regression, making this method conceptually closer to Enet than to XGBoost. In other words, the question is not to choose between statistical and machine learning but the best method according to the context.

5.3.3 Feature selection

We also evaluated how expert-driven feature selection compares to model-based selection using all available features. The expert-selected set included hospitalizations, positive RT-PCR cases, the proportion of positive RT-PCRs among

individuals over 60, IPTCC, the cumulative number of first-dose vaccinations, and emergency department visits with “COVID-19” mentioned in the EHR. These variables were chosen by an epidemiologist based on prior knowledge before fitting the models.

As Table 2 shows, differences in MAE were under 0.5 hospitalizations across all algorithms. Enet was unaffected. RC showed a slight improvement with the expert-selected features, possibly due to the advantage of starting with a smaller, more focused input set before reservoir-based dimensional expansion. XGBoost performed slightly worse with the expert-selected features.

While the performance differences are modest, expert-driven feature selection may offer benefits from an implementation science perspective, such as improved interpretability and ease of adoption, as discussed in Section 7.

6 Key Point 4: Uncertainty quantification

Uncertainty quantification is essential for informed decision-making, especially in COVID-19 forecasting where epidemiological dynamics change rapidly. Estimating prediction intervals is challenging due to temporal correlations.

Initially, we applied a simple $\pm 40\%$ range around predictions, which yielded only 51% empirical coverage—insufficient for epidemic forecasting (11). We refined this approach, denoted as Ad hoc in Figure 3, by expanding intervals by $\pm x\%$, calibrated to achieve 95% coverage using data from 2020-09-02 to 2021-03-01 as a calibration set.

To further improve uncertainty estimation, we considered model-agnostic post hoc methods that leverage past residuals. In our setup, residuals from 2020-09-02 to 2021-03-01 form an initial calibration set, updated daily as new residuals become available, and are used to evaluate three additional approaches.

Two baseline methods are shown in Figure 3: *Residuals*, which assumes normality to compute 95% intervals, and *Bootstrap (BS)*, which samples 1,000 past residuals to construct empirical 95% quantiles (52). Both methods adapt slowly, as older residuals are weighted equally with more recent ones.

Conformal prediction (CP) addresses this limitation. Adaptive Conformal Inference (ACI) adjusts interval widths daily based on prior coverage, controlled by a hyperparameter γ (54). AgACI extends this approach by automatically tuning γ via online expert aggregation (53), implemented in the `AdaptiveConformal` R package (55).

Figure 3 summarizes the performance of these methods; additional details are in Supplementary 10.2. All approaches except the Ad hoc method reached the nominal 95% coverage, though some slightly over-covered. AgACI achieved the narrowest intervals, followed closely by BS, while Residuals underperformed, likely due to the normality assumption. The Ad hoc strategy failed to reach 95% coverage.

Looking ahead, BS and CP methods appear to be the most promising for future epidemic forecasting. CP is theoretically well-suited to handle abrupt distributional shifts, due to its online adaptive nature. This is a particularly

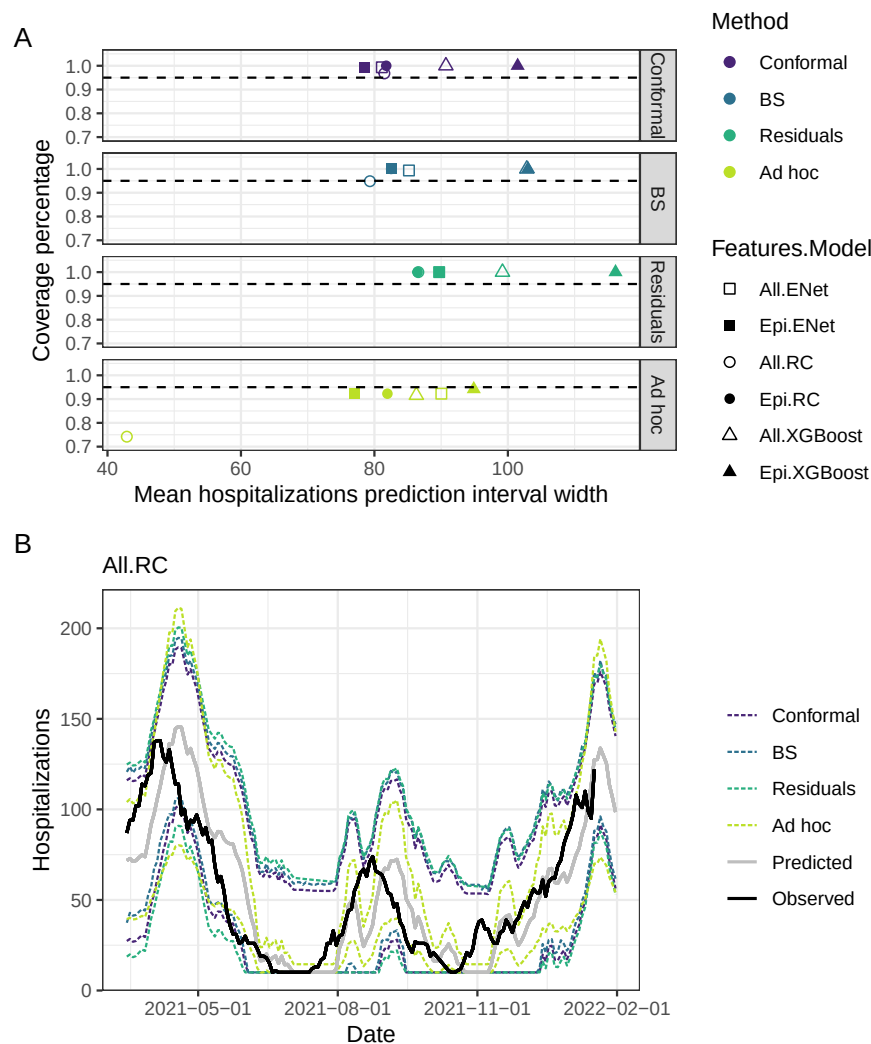


Figure 3: Estimating model uncertainty. A) Coverage percentage and width of 95% prediction intervals (PI); the dotted line marks the target coverage. The best method achieves coverage of 95% with the narrowest interval width. B) Prediction intervals for RC predictions using all features and different estimation methods. *Residuals* and *BS* estimate PIs from the distribution of past residuals (52). *Conformal* applies an adaptive method to iteratively widen or tighten the PIs (53). The *Ad hoc* method adds $\pm x\%$ to predictions based on residuals observed before 2021-03-01 (11).

valuable property in the context of rapidly evolving epidemics. However, in practice, our empirical results showed little difference in performance between CP and BS methods.

7 Key Point 5: Considering Real-World Algorithm Implementation

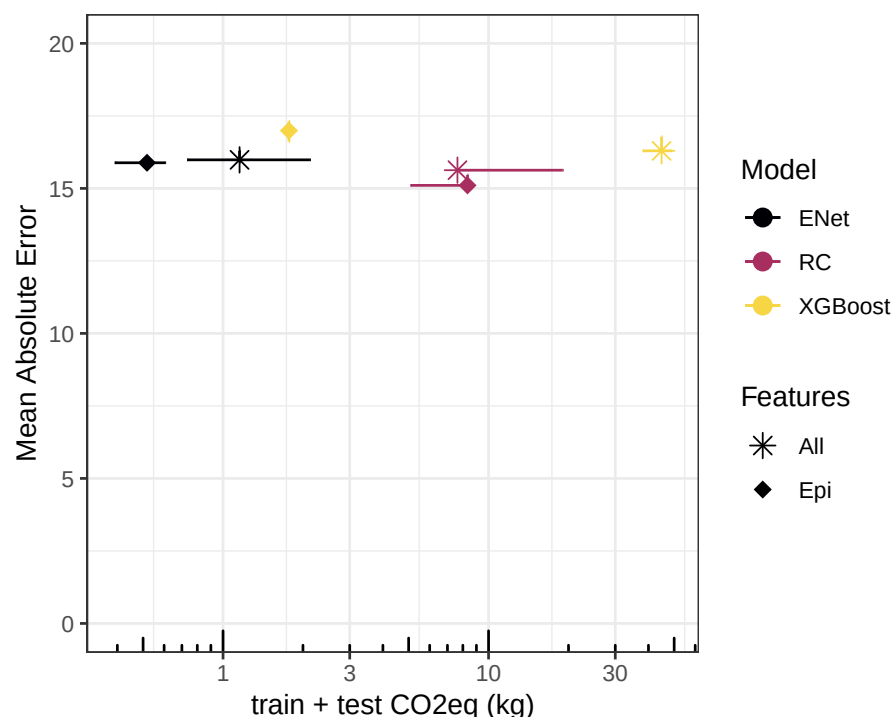


Figure 4: Carbon footprint and performance. Models were initiated on 2021-03-01 with monthly hyperparameter updates. Dots correspond to the median of three replications of the analysis; bars correspond to the minimum and maximum of the three replications.

While RC achieved the highest forecasting performance, its advantage over ENet was modest, highlighting the need to consider criteria beyond predictive accuracy. Using the Consolidated Framework for Implementation Research (CFIR) (56), we assess model deployment across five implementation domains, emphasizing the Innovation domain.

Innovation	ENet	XGBoost	RC
Relative Advantage	⚖️ Slightly lower performance	✗ Lower performance	✓ Best forecasting accuracy
Complexity	✓ Simple and familiar	⚖️ Standard libraries available	✗ Less common, more complex due to genetic tuning
Cost	✓ Low computational and human resource cost	⚖️ Computational cost depends on feature set	✗ Moderate computational cost, higher human effort
Design Quality & Packaging	✓ Well-supported and easy to integrate into standard pipelines		✗ Requires custom integration and maintenance for genetic optimization
Trialability	✓ All models can be readily tested retrospectively with standard cross-validation frameworks		

Table 3: Comparison of ENet, XGBoost, and RC for the Innovation domain of the CFIR framework.

7.1 Innovation

The Innovation domain in CFIR focuses on characteristics of the intervention, including Relative Advantage, Complexity, Cost, Design Quality and Packaging, and Trialability. Table 3 summarizes the three forecasting models on these dimensions.

Relative Advantage. RC achieved the lowest mean absolute error, but its advantage over ENet was modest, raising the question of whether the performance gain justifies the added complexity.

Complexity. ENet is the simplest and widely familiar. XGBoost is also well-known with robust libraries. RC, particularly with genetic algorithm tuning, is technically complex and less familiar to practitioners.

Cost. ENet has a clear advantage in terms of both computational and human resource costs, which remain low even at scale. XGBoost’s computational cost varies depending on the number of features and parameter settings, but its practical cost remains moderate. In contrast, RC with genetic optimization introduces moderate computational demands and requires additional human expertise for setup and maintenance.

Design Quality and Packaging. ENet and XGBoost are well-supported within existing data science ecosystems, offering straightforward implementation and mature libraries in both `Python` and `R`. RC, particularly when combined with genetic algorithms, is less seamlessly integrated. However, the `ReservoirPy` module provides a flexible Python implementation, and we recently developed the `reservoirnet` package to extend this accessibility to R users (14, 57).

Trialability. All three models have high trialability, as they can be rigorously tested on retrospective data before daily deployment. We implemented an accumulative validation framework (58), where models were retrained daily on all available data to mimic real-time conditions, with hyperparameters re-optimized monthly to reflect realistic resource constraints. Full details are provided in Section 10.1.2 and Ferté et al. (12).

Overall, this comparison highlights a trade-off between accuracy and ease of implementation. RC delivers highest performance, while ENet’s simplicity, low cost, and familiarity make it a strong alternative. In our setting, XGBoost offered no clear performance advantage to justify its added complexity.

7.2 Inner Setting

Implementation of forecasting models was enabled by a well-established hospital data warehouse and an existing daily dashboard reporting key indicators such as hospitalizations, confirmed cases, and ICU admissions, providing a strong foundation for integration (11).

Computing capacity posed little barrier for lightweight models like ENet, though more resource-intensive methods such as RC with genetic algorithms could have been constrained.

Strong collaboration between the epidemiology–biostatistics and medical informatics teams, built through prior joint projects, fostered trust and a shared problem-solving culture, which supported timely implementation (59).

7.3 Outer Setting

The SARS-CoV-2 pandemic was the main external driver, creating urgent demand for timely data to guide crisis response and mobilizing healthcare professionals.

While no direct environmental constraints affected our project, awareness is growing around the carbon footprint of data-intensive methods. As shown in Figure 4, ENet had the lowest emissions, XGBoost’s footprint increased with feature set size, and RC with genetic algorithms was most energy-intensive. The burden also depends on local energy mixes: France’s electricity (51.28 gCO₂/kWh) is far less carbon-intensive than Germany’s (338.66) or the U.S.’s (423.94) (60).

7.4 Individuals

Implementation relied on the knowledge, attitudes, and involvement of key individuals. Interdisciplinary collaboration across informatics, epidemiology, biostatistics, and machine learning was essential (11–14). Strong leadership and early engagement of the informatics team improved understanding of both data infrastructure and model assumptions.

Close collaboration between those developing, deploying, and interpreting the model helped ensure that forecasts were used appropriately. As shown in Figures 2 and 3, predictions are inherently uncertain and must be contextualized with other epidemiological indicators.

Finally, introducing more complex methods like RC highlights a potential barrier: these models are less familiar compared to linear regression. Addressing this gap is part of the motivation for this work. By documenting our experience, we aim to support knowledge sharing and reduce this barrier for future implementations.

7.5 Implementation process

As noted in Section 7.4, effective forecasting in routine practice requires close collaboration across multiple teams. This interdisciplinary work helps ensure that implementation is aligned with the realities of data acquisition, local constraints, and available infrastructure.

A staged approach is pragmatic: begin with simple models and selected features, then increase complexity as capacity grows. Enet with expert-selected variables provided a strong baseline (Figure 4), while RC offered modest improvements. This is further supported by Table 1, which shows that when limited information was available, none of the models substantially outperformed the baseline.

Finally, timely engagement of decision-makers through clear communication remains essential, as further discussed in Section 8.

8 Key Point 6: Communicate Effectively with Policymakers

Even the most accurate forecasting algorithm is ineffective if it does not translate into actionable decisions. Therefore, model predictions should be designed to support policymakers in making informed choices. However, given the inherent uncertainties surrounding SARS-CoV-2 dynamics, it remains unclear how best to communicate these predictions.

In our setting, access to model outputs was restricted to public health experts, who interpreted the predictions alongside other epidemiological indicators. These interpretations were then discussed at each hospital Covid-19 task force meeting, together with hospital indicators. This decision was primarily

driven by two factors: (1) the moderate predictive performance of the models, which limited their standalone reliability, and (2) the rapid evolution of the epidemic, which was frequently influenced by external interventions such as lockdowns and curfews. Given these challenges, presenting raw model outputs to decision-makers without adequate context risked misinterpretation. Alternative communication strategies could be considered, but it is likely that model predictions should always be contextualized before being shared with policymakers to ensure proper understanding.

9 Discussion

Summary Box

- **Integrate multiple data sources:** The most accurate prediction models combine various data modalities. For complex diseases, a multimodal representation is crucial.
- **Leverage real-time data:** Access to current information significantly enhances forecasting accuracy. Hospital data warehouses provide a powerful means of obtaining near real-time insights.
- **Establish a baseline model:** Predictive performance should always be benchmarked against a simple, transparent baseline model.
- **Choose evaluation metrics carefully:** Relative metrics (e.g., Mean Relative Error) can obscure significant errors when case numbers are high. Prioritize absolute metrics.
- **Start with simple, easy-to-implement models:** Linear regression, combined with expert-driven feature selection, offers a practical, maintainable starting point.
- **Quantify uncertainty:** Predictions should include prediction intervals to inform decision-makers about the associated uncertainty. Bootstrapping residuals and conformal prediction are promising methods to address this challenge.
- **Evaluate advanced methods cautiously:** Complex models (e.g., RC) may improve accuracy but require careful consideration of implementation and maintenance costs.
- **Contextualize forecasts for decision-making:** Predictions should be assessed alongside epidemiological indicators and baseline comparisons to ensure relevance and reliability.
- **Ensure real-world impact:** A predictive model must support actionable public health interventions to be considered effective.

A summary of the main findings is provided in Box 9. This study of epidemic forecasting at BUH underscores the value of high-quality, real-time data from Hospital Data Warehouses and of a transparent baseline model to assess more complex approaches. Linear regression with expert-selected features offered a strong initial choice due to its simplicity, interpretability, and low maintenance, while models like RC with genetic algorithms improved accuracy but demanded greater computational resources, expertise, and upkeep. Forecasts should always be contextualized with epidemiological indicators and validated against baselines before informing policy, particularly in dynamic situations such as

COVID-19.

Our findings align with recent work emphasizing diverse, real-time data integration and strong collaboration between modelers and decision-makers. Nunes et al. (10) advocate multimodal inputs such as wastewater, mobility, and social media data, highlighting challenges in trend shifts and the potential of ensemble methods. Endres-Dighe et al. (61) emphasize practical issues including computational demands, automation, and structured communication via dashboards and protocols. Across approaches, shared challenges were responsiveness to evolving data, automation, and stakeholder engagement.

Future work should improve adaptation to emerging epidemics. The cumulative validation framework used here weights all data points equally, which diminishes the influence of recent observations as the dataset grows. Approaches proposed by Montavon et al. (62) and Lukoševičius et al. (58) suggest emphasizing recent or critical points, such as trend shifts.

Finally, black-box models like RC and XGBoost may reduce trust; explainable AI methods (e.g., Shapley values) can clarify which features drive predictions (63, 64).

Effective epidemic forecasting requires integrating diverse real-time data, grounding predictions in simple baselines, and choosing appropriate evaluation metrics. Starting with simple models enables rapid deployment while providing time to develop more complex approaches. Above all, forecasts must be contextualized and aligned with public health decision-making to achieve real-world impact.

Acknowledgements

Experiments presented in this paper were conducted using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Université de Bordeaux, Bordeaux INP and Conseil Régional d'Aquitaine (see <https://www.plafrim.fr>), as well as by the MCIA (Mésocentre de Calcul Intensif Aquitain).

Funding

This study was carried out in the framework of the University of Bordeaux's France 2030 program / RRI PHDS.

Conflicts of interest

Authors declare no conflicts of interest.

Data availability

The data supporting this study is publicly available (65) . The code used in this study is also publicly available (66) .

CRedit

RT, BH, and VJ contributed to conceptualization. VJ and RG were responsible for data curation. TF performed the formal analysis. Methodology was developed by TF, VT, XH, PL, DD, RG, VJ, BH, and RT. Project administration was carried out by RT and VJ. TF and DD developed the software, while RT, BH, VJ, and XH provided supervision. Visualization was conducted by TF. The original draft was written by TF, VT, VJ, BH, and RT, and XH, PL, DD, and RG contributed to writing—review and editing.

References

- (1) Fan Wu, Su Zhao, Bin Yu, et al. “A new coronavirus associated with human respiratory disease in China”. en. In: *Nature* 579.7798 (Mar. 2020). Publisher: Nature Publishing Group, pp. 265–269. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2008-3.
- (2) World Health Organization. *COVID-19 – Chronologie de l’action de l’OMS*. fr. Apr. 2020.
- (3) CDC. *CDC Museum COVID-19 Timeline*. en-us. July 2024.
- (4) Juan Li, Shengjie Lai, George F. Gao, et al. “The emergence, genomic diversity and global spread of SARS-CoV-2”. en. In: *Nature* 600.7889 (Dec. 2021). Publisher: Nature Publishing Group, pp. 408–418. ISSN: 1476-4687. DOI: 10.1038/s41586-021-04188-6.
- (5) Thomas Hale, Noam Angrist, Rafael Goldszmidt, et al. “A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)”. en. In: *Nature Human Behaviour* 5.4 (Apr. 2021). Publisher: Nature Publishing Group, pp. 529–538. ISSN: 2397-3374. DOI: 10.1038/s41562-021-01079-8.
- (6) Iris Ganser, David L. Buckeridge, Jane Heffernan, et al. “Estimating the population effectiveness of interventions against COVID-19 in France: A modelling study”. In: *Epidemics* 46 (Mar. 2024), p. 100744. ISSN: 1755-4365. DOI: 10.1016/j.epidem.2024.100744.
- (7) Costase Ndayishimiye, Christoph Sowada, Patrycja Dyjach, et al. “Associations between the COVID-19 Pandemic and Hospital Infrastructure Adaptation and Planning—A Scoping Review”. In: *International Journal of Environmental Research and Public Health* 19.13 (July 2022), p. 8195. ISSN: 1661-7827. DOI: 10.3390/ijerph19138195.

- (8) Taulant Muka, Joshua J. X. Li, Sahar J. Farahani, et al. “An umbrella review of systematic reviews on the impact of the COVID-19 pandemic on cancer prevention and management, and patient needs”. eng. In: *eLife* 12 (Apr. 2023), e85679. ISSN: 2050-084X. DOI: 10.7554/eLife.85679.
- (9) Carmela Comito and Clara Pizzuti. “Artificial intelligence for forecasting and diagnosing COVID-19 pandemic: A focused review”. In: *Artificial Intelligence in Medicine* 128 (June 2022), p. 102286. ISSN: 0933-3657. DOI: 10.1016/j.artmed.2022.102286.
- (10) Marta C. Nunes, Edward Thommes, Holger Fröhlich, et al. “Redefining pandemic preparedness: Multidisciplinary insights from the CERP modelling workshop in infectious diseases, workshop report”. In: *Infectious Disease Modelling* 9.2 (Feb. 2024), pp. 501–518. ISSN: 2468-2152. DOI: 10.1016/j.idm.2024.02.008.
- (11) Thomas Ferté, Vianney Jouhet, Romain Griffier, et al. “The benefit of augmenting open data with clinical data-warehouse EHR for forecasting SARS-CoV-2 hospitalizations in Bordeaux area, France”. eng. In: *JAMIA open* 5.4 (Dec. 2022), ooac086. ISSN: 2574-2531. DOI: 10.1093/jamiaopen/ooac086.
- (12) Thomas Ferté, Dan Dutartre, Boris P. Hejblum, et al. “Reservoir Computing for Short High-Dimensional Time Series: an Application to SARS-CoV-2 Hospitalization Forecast”. en. In: *Proceedings of the 41st International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, July 2024, pp. 13570–13591.
- (13) Thomas Ferté, Dan Dutartre, Boris P. Hejblum, et al. “Optimizing Reservoir Computing with Genetic Algorithm for High-Dimensional SARS-CoV-2 Hospitalization Forecasting: Impacts of Genetic Algorithm Hyperparameters on Feature Selection and Reservoir Computing Hyperparameter Tuning”. In: *Lecture Notes in Computer Science*. Lecture Notes in Computer Science. Pierrick Legrand. Bordeaux, France, Oct. 2024.
- (14) Thomas Ferté, Kalidou Ba, Dutartre Dan, et al. “Reservoir Computing in R: a Tutorial for Using reservoirnet to Predict Complex Time-Series”. en. In: *Computo* (May 2025).
- (15) Etalab. *Les données relatives au COVID-19 en France - data.gouv.fr*. fr. 2020.
- (16) Adam Smith, Neal Lott, and Russ Vose. “The Integrated Surface Database: Recent Developments and Partnerships”. EN. In: *Bulletin of the American Meteorological Society* 92.6 (June 2011). Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, pp. 704–708. DOI: 10.1175/2011BAMS3015.1.

- (17) Thomas Ferté, Vianney Jouhet, Romain Griffier, et al. *The benefit of augmenting open data with clinical data-warehouse EHR for forecasting SARS-CoV-2 hospitalizations in Bordeaux area, France*. en. Artwork Size: 153575 bytes Pages: 153575 bytes. Jan. 2023. DOI: 10.5061/DRYAD.HHMQNKX.
- (18) Farrukh Saleem, Abdullah Saad Al-Malaise Al-Ghamdi, Madini O. Alasafi, et al. “Machine Learning, Deep Learning, and Mathematical Models to Analyze Forecasting and Epidemiology of COVID-19: A Systematic Literature Review”. eng. In: *International Journal of Environmental Research and Public Health* 19.9 (Apr. 2022), p. 5099. ISSN: 1660-4601. DOI: 10.3390/ijerph19095099.
- (19) Herbert Jaeger. “The” echo state” approach to analysing and training recurrent neural networks-with an erratum note”. In: *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report* 148 (Jan. 2001).
- (20) Wolfgang Maass, Thomas Natschläger, and Henry Markram. “Real-time computing without stable states: a new framework for neural computation based on perturbations”. eng. In: *Neural Computation* 14.11 (Nov. 2002), pp. 2531–2560. ISSN: 0899-7667. DOI: 10.1162/089976602760407955.
- (21) Mantas Lukoševičius and Herbert Jaeger. “Reservoir computing approaches to recurrent neural network training”. In: *Computer Science Review* 3.3 (Aug. 2009), pp. 127–149. ISSN: 1574-0137. DOI: 10.1016/j.cosrev.2009.03.005.
- (22) P. R. Vlachas, J. Pathak, B. R. Hunt, et al. “Backpropagation algorithms and Reservoir Computing in Recurrent Neural Networks for the forecasting of complex spatiotemporal dynamics”. In: *Neural Networks* 126 (June 2020), pp. 191–217. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2020.02.016.
- (23) Xavier Hinaut and Peter Ford Dominey. “Real-Time Parallel Processing of Grammatical Structure in the Fronto-Striatal System: A Recurrent Network Simulation Study Using Reservoir Computing”. en. In: *PLOS ONE* 8.2 (Feb. 2013). Publisher: Public Library of Science, e52946. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0052946.
- (24) Subrata Ghosh, Abhishek Senapati, Arindam Mishra, et al. “Reservoir computing on epidemic spreading: A case study on COVID-19 cases”. en. In: *Physical Review E* 104.1 (July 2021), p. 014308. ISSN: 2470-0045, 2470-0053. DOI: 10.1103/PhysRevE.104.014308.
- (25) Arnob Ray, Tanujit Chakraborty, and Dibakar Ghosh. “Optimized ensemble deep learning framework for scalable forecasting of dynamics containing extreme events”. eng. In: *Chaos (Woodbury, N.Y.)* 31.11 (Nov. 2021), p. 111105. ISSN: 1089-7682. DOI: 10.1063/5.0074213.

- (26) Bocheng Liu, Yiyuan Xie, Weichen Liu, et al. “Nanophotonic reservoir computing for COVID-19 pandemic forecasting”. eng. In: *Nonlinear Dynamics* 111.7 (2023), pp. 6895–6914. ISSN: 0924-090X. DOI: 10.1007/s11071-022-08190-z.
- (27) Min Yan, Can Huang, Peter Bienstman, et al. “Emerging opportunities and challenges for the future of reservoir computing”. In: *Nature Communications* 15.1 (2024), p. 2056.
- (28) Heng Zhang and Danilo Vasconcellos Vargas. “A Survey on Reservoir Computing and its Interdisciplinary Applications Beyond Traditional Machine Learning”. In: *IEEE Access* 11 (2023). Conference Name: IEEE Access, pp. 81033–81070. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3299296.
- (29) Estee Y. Cramer, Evan L. Ray, Velma K. Lopez, et al. “Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States”. In: *Proceedings of the National Academy of Sciences* 119.15 (Apr. 2022). Publisher: Proceedings of the National Academy of Sciences, e2113561119. DOI: 10.1073/pnas.2113561119.
- (30) Poulami Majumder and Partha Pratim Ray. “A systematic review and meta-analysis on correlation of weather with COVID-19”. eng. In: *Scientific Reports* 11.1 (May 2021), p. 10746. ISSN: 2045-2322. DOI: 10.1038/s41598-021-90300-9.
- (31) Alix Roumagnac, Eurico de Carvalho Filho, Raphaël Bertrand, et al. “Étude de l’influence potentielle de l’humidité et de la température dans la propagation de la pandémie COVID-19”. In: *Médecine De Catastrophe, Urgences Collectives* 5.1 (Mar. 2021), pp. 87–102. ISSN: 1279-8479. DOI: 10.1016/j.pxur.2021.01.002.
- (32) Zella King, Joseph Farrington, Martin Utley, et al. “Machine learning for real-time aggregated prediction of hospital admission for emergency patients”. eng. In: *NPJ digital medicine* 5.1 (July 2022), p. 104. ISSN: 2398-6352. DOI: 10.1038/s41746-022-00649-y.
- (33) Riza C. Li, Cecelia K. Harrison, Claudine T. Jurkovitz, et al. “Prediction model for COVID-19 patient visits in the ambulatory setting”. eng. In: *Research Square* (Mar. 2021), rs.3.rs-177379. ISSN: 2693-5015. DOI: 10.21203/rs.3.rs-177379/v1.
- (34) Juliette Paireau, Alessio Andronico, Nathanaël Hozé, et al. “An ensemble model based on early predictors to forecast COVID-19 health care demand in France”. en. In: *Proceedings of the National Academy of Sciences* 119.18 (May 2022), e2103302119. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2103302119.
- (35) Cornelia Ilin, Sébastien Annan-Phan, Xiao Hui Tai, et al. “Public mobility data enables COVID-19 forecasting and management at local and global scales”. eng. In: *Scientific Reports* 11.1 (June 2021), p. 13531. ISSN: 2045-2322. DOI: 10.1038/s41598-021-92892-8.

- (36) Christian Selinger, Marc Choisy, and Samuel Alizon. “Predicting COVID-19 incidence in French hospitals using human contact network analytics”. eng. In: *International journal of infectious diseases: IJID: official publication of the International Society for Infectious Diseases* 111 (Oct. 2021), pp. 100–107. ISSN: 1878-3511. DOI: 10.1016/j.ijid.2021.08.029.
- (37) R. Thiébaut and F. Thiessard. “Public Health and Epidemiology Informatics”. In: *Yearbook of Medical Informatics* 26.1 (Aug. 2017), pp. 248–251. ISSN: 0943-4747. DOI: 10.15265/IY-2017-036.
- (38) Tobias Saegner and Donatas Austys. “Forecasting and Surveillance of COVID-19 Spread Using Google Trends: Literature Review”. In: *International Journal of Environmental Research and Public Health* 19.19 (Sept. 2022), p. 12394. ISSN: 1661-7827. DOI: 10.3390/ijerph191912394.
- (39) Berlanty A Zayed, Ahmed M Talaia, Mohamed A Gaaboobah, et al. “Google Trends as a predictive tool in the era of COVID-19: a scoping review”. In: *Postgraduate Medical Journal* 99.1175 (Sept. 2023), pp. 962–975. ISSN: 0032-5473. DOI: 10.1093/postmj/qgad012.
- (40) Surabhi Singh, Amina Ismail Ahmed, Sumayya Almansoori, et al. “A narrative review of wastewater surveillance: pathogens of concern, applications, detection methods, and challenges”. English. In: *Frontiers in Public Health* 12 (July 2024). Publisher: Frontiers. ISSN: 2296-2565. DOI: 10.3389/fpubh.2024.1445961.
- (41) Chen Chen, Yunfan Wang, Gursharn Kaur, et al. “Wastewater-based epidemiology for COVID-19 surveillance and beyond: A survey”. In: *Epidemics* 49 (Dec. 2024), p. 100793. ISSN: 1755-4365. DOI: 10.1016/j.epidem.2024.100793.
- (42) Fatemeh Torabi, Guangquan Li, Callum Mole, et al. “Wastewater-based surveillance models for COVID-19: A focused review on spatio-temporal models”. eng. In: *Heliyon* 9.11 (Nov. 2023), e21734. ISSN: 2405-8440. DOI: 10.1016/j.heliyon.2023.e21734.
- (43) Michael D. Parkins, Bonita E. Lee, Nicole Acosta, et al. “Wastewater-based surveillance as a tool for public health action: SARS-CoV-2 and beyond”. eng. In: *Clinical Microbiology Reviews* 37.1 (Mar. 2024), e0010322. ISSN: 1098-6618. DOI: 10.1128/cmr.00103-22.
- (44) J. Bracher, D. Wolfram, J. Deuschel, et al. “A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave”. en. In: *Nature Communications* 12.1 (Aug. 2021). Publisher: Nature Publishing Group, p. 5173. ISSN: 2041-1723. DOI: 10.1038/s41467-021-25207-0.
- (45) Nanning Zheng, Shaoyi Du, Jianji Wang, et al. “Predicting COVID-19 in China Using Hybrid AI Model”. In: *IEEE Transactions on Cybernetics* 50.7 (July 2020), pp. 2891–2904. ISSN: 2168-2275. DOI: 10.1109/TCYB.2020.2990162.

- (46) Philip Gerlee, Anna Jöud, Armin Spreco, et al. “Computational models predicting the early development of the COVID-19 pandemic in Sweden: systematic review, data synthesis, and secondary validation of accuracy”. eng. In: *Scientific Reports* 12.1 (Aug. 2022), p. 13256. ISSN: 2045-2322. DOI: 10.1038/s41598-022-16159-6.
- (47) Jonas Botz, Diego Valderrama, Jannis Guski, et al. “A dynamic ensemble model for short-term forecasting in pandemic situations”. In: *PLOS Global Public Health* 4.8 (Aug. 2024), e0003058. ISSN: 2767-3375. DOI: 10.1371/journal.pgph.0003058.
- (48) Samuel G. Finlayson, Andrew L. Beam, and Maarten van Smeden. “Machine Learning and Statistics in Clinical Research Articles—Moving Past the False Dichotomy”. In: *JAMA Pediatrics* 177.5 (May 2023), pp. 448–450. ISSN: 2168-6203. DOI: 10.1001/jamapediatrics.2023.0034.
- (49) Peter C Austin, Frank E Harrell Jr, and Ewout W Steyerberg. “Predictive performance of machine and statistical learning methods: Impact of data-generating processes on external validity in the “large N, small p” setting”. EN. In: *Statistical Methods in Medical Research* 30.6 (June 2021). Publisher: SAGE Publications Ltd STM, pp. 1465–1483. ISSN: 0962-2802. DOI: 10.1177/09622802211002867.
- (50) Andrew L. Beam and Isaac S. Kohane. “Big Data and Machine Learning in Health Care”. In: *JAMA* 319.13 (Apr. 2018), pp. 1317–1318. ISSN: 0098-7484. DOI: 10.1001/jama.2017.18391.
- (51) Ravikiran Keshavamurthy, Samuel Dixon, Karl T. Pazdernik, et al. “Predicting infectious disease for biopreparedness and response: A systematic review of machine learning and deep learning approaches”. eng. In: *One Health (Amsterdam, Netherlands)* 15 (Dec. 2022), p. 100439. ISSN: 2352-7714. DOI: 10.1016/j.onehlt.2022.100439.
- (52) Rob J Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice (2nd ed)*. OTexts: Melbourne, Australia, May 2025.
- (53) Margaux Zaffran, Olivier Feron, Yannig Goude, et al. “Adaptive Conformal Predictions for Time Series”. en. In: *Proceedings of the 39th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, June 2022, pp. 25834–25866.
- (54) Isaac Gibbs and Emmanuel Candes. “Adaptive Conformal Inference Under Distribution Shift”. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 1660–1672.
- (55) Herbert Susmann, Antoine Chambaz, and Julie Josse. “Adaptive Conformal: An ‘R’ Package for Adaptive Conformal Inference”. en. In: *Computo* (July 2024). Publisher: Société Française de Statistique. ISSN: 2824-7795. DOI: 10.57750/edan-5f53.

- (56) Laura J. Damschroder, Caitlin M. Reardon, Marilla A. Opra Widerquist, et al. “The updated Consolidated Framework for Implementation Research based on user feedback”. In: *Implementation Science* 17.1 (Oct. 2022), p. 75. ISSN: 1748-5908. DOI: 10.1186/s13012-022-01245-0.
- (57) Thomas Ferte, Kalidou Ba, Nathan Trouvain, et al. *reservoirnet: Reservoir Computing and Echo State Networks*. June 2025.
- (58) Mantas Lukoševičius and Arnas Uselis. “Efficient implementations of echo state network cross-validation”. en. In: *Cognitive Computation* 15.5 (Sept. 2023). arXiv:2006.11282 [cs], pp. 1470–1484. ISSN: 1866-9956, 1866-9964. DOI: 10.1007/s12559-021-09849-2.
- (59) Thomas Ferté, Sébastien Cossin, Thierry Schaefferbeke, et al. “Automatic phenotyping of electronic health record: PheVis algorithm”. en. In: *Journal of Biomedical Informatics* 117 (May 2021), p. 103746. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2021.103746.
- (60) Loïc Lannelongue, Jason Grealey, and Michael Inouye. “Green Algorithms: Quantifying the Carbon Footprint of Computation”. en. In: *Advanced Science* 8.12 (2021), p. 2100707. ISSN: 2198-3844. DOI: 10.1002/adv.202100707.
- (61) Stacy Endres-Dighe, Kasey Jones, Emily Hadley, et al. “Lessons learned from the rapid development of a statewide simulation model for predicting COVID-19’s impact on healthcare resources and capacity”. eng. In: *PloS One* 16.11 (2021), e0260310. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0260310.
- (62) Mantas Lukoševičius. “A Practical Guide to Applying Echo State Networks”. en. In: *Neural Networks: Tricks of the Trade*. Ed. by Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller. Vol. 7700. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 659–686. ISBN: 978-3-642-35288-1. DOI: 10.1007/978-3-642-35289-8_36.
- (63) Jessica Caterson, Alexandra Lewin, and Elizabeth Williamson. “The application of explainable artificial intelligence (XAI) in electronic health record research: A scoping review”. In: *Digital Health* 10 (Oct. 2024), p. 20552076241272657. ISSN: 2055-2076. DOI: 10.1177/20552076241272657.
- (64) Kjersti Aas, Martin Jullum, and Anders Løland. “Explaining individual predictions when features are dependent: More accurate approximations to Shapley values”. In: *Artificial Intelligence* 298 (Sept. 2021), p. 103502. ISSN: 0004-3702. DOI: 10.1016/j.artint.2021.103502.
- (65) thomasferte. *thomasferte/Ferte2024ICML_HighDimensionReservoir: Final version ICML*. Version v1.0.0. Oct. 2025. DOI: 10.5281/zenodo.17295479.
- (66) thomasferte. *thomasferte/EpidemioForecast: Submission ready*. Version v1.0.0. Oct. 2025. DOI: 10.5281/zenodo.17295500.

- (67) David Carslaw. *worldmet: Import Surface Meteorological Data from NOAA Integrated Surface Database (ISD)*. June 2023.
- (68) Li Yang and Abdallah Shami. “On hyperparameter optimization of machine learning algorithms: Theory and practice”. In: *Neurocomputing* 415 (Nov. 2020), pp. 295–316. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2020.07.061.
- (69) Nathan Trouvain and Xavier Hinaut. *reservoirpy: A Simple and Flexible Reservoir Computing Tool in Python*. en. June 2022.

10 Supplementary

10.1 Supplementary methods

To illustrate various aspects of epidemic forecasting implementation, we use the example of SARS-CoV-2 forecasting at BUH, employing data and methods similar to those in Ferté et al. (12). The following sections will detail the data, evaluation framework, modeling approaches, hyperparameter optimization and feature selection methods.

10.1.1 Data

The objective of this study was to forecast the number of SARS-CoV-2 hospitalized patients at 14 days (h_{t+14}) at BUH. To achieve this, we integrated data from multiple sources: department-level COVID-19 statistics from Santé Publique France (15), department-level weather data from the National Oceanic and Atmospheric Administration (NOAA) (16), and hospital-level COVID-19 data from BUH’s Electronic Health Records (EHR) (17). The dataset covered the period from May 16, 2020, to January 17, 2022, resulting in a high-dimensional problem with 409 predictors over 586 days, where each observation corresponds to a single day.

For preprocessing, we computed the first and second derivatives over the previous seven days to enrich the model with temporal dynamics. Additionally, we applied local polynomial regression smoothing with a 21-day span to reduce daily noise variations, following recommendations from Ferté et al. (11). Finally, all features were scaled between -1 and 1 using maximum absolute scaling, where each value was divided by the maximum absolute value of the respective feature.

All datasets used in this study are publicly available. Weather data can be accessed via Smith et al. (16) through the R package *worldmet* (67), and vaccine data can be downloaded from Etalab (15) here. EHR data are available on Dryad (17), with privacy protections in place. For that purpose, EHR values below 10 patients were obfuscated to 0. To account for this, when evaluating model performance, outcomes, forecasts, and hospitalization values of 0 were replaced with 10. The full dataset and code are accessible via a GitHub repository https://github.com/thomasferte/-Ferte2024ICML_HighDimensionReservoir.

10.1.2 Evaluation framework

Model performance was evaluated using two different implementation periods: an "early" model starting on September 2, 2021, and a "late" model starting on March 1, 2021. Data prior to these start dates (i.e., before September 2, 2021, for the early model and before March 1, 2021, for the late model) were used solely for identifying relevant hyperparameters. The remaining data were used to assess model performance, with hyperparameters being updated (or left unchanged) on a monthly basis.

These two periods were chosen to examine how model performance differs when trained with limited historical data versus when more information is available. The early period represents a scenario where forecasting is more challenging due to limited prior observations, while the late period illustrates what can be achieved with a longer historical context. Our hypothesis was that, with limited data, simpler models incorporating expert-driven feature selection would outperform more complex approaches. However, as more information becomes available, more sophisticated models might demonstrate superior performance.

The primary evaluation metric was MAE, while secondary metrics included Median Relative Error (MRE), Mean Absolute Error to Baseline (MAEB), and Median Relative Error to Baseline (MREB). These metrics were defined as follows:

- $MAE = mean(|\widehat{h_{t+14}} - h_{t+14}|).$
- $MRE = median(|\frac{\widehat{h_{t+14}} - h_{t+14}}{h_{t+14}}|).$
- $MAEB = mean(|\widehat{h_{t+14}} - h_{t+14}| - |h_t - h_{t+14}|).$
- $MREB = median(|\frac{\widehat{h_{t+14}} - h_{t+14}}{h_t - h_{t+14}}|)$

The relevance of these different metrics will be discussed in the following sections, particularly by contrasting absolute and relative errors and emphasizing the importance of baseline-derived metrics.

To limit computational costs, models were trained and evaluated every other day rather than on a daily basis using an accumulative validation scheme (58). This scheme, illustrated in Figure S1, updates hyperparameters monthly based on all data available up to the beginning of each month, as described in Section 10.1.4. The optimal set of hyperparameters is selected and used to forecast the following month. At the start of the next month, new candidate hyperparameters are evaluated on all available data up to that point, and the best set is again chosen for forecasting. This cycle continues until the end of the dataset.

10.1.3 Models and Implementation

We compared three different models: ENet, XGBoost, and RC.

- ENet was selected as a representative of statistical models. It is a simple, interpretable model that is easy to use and implement.

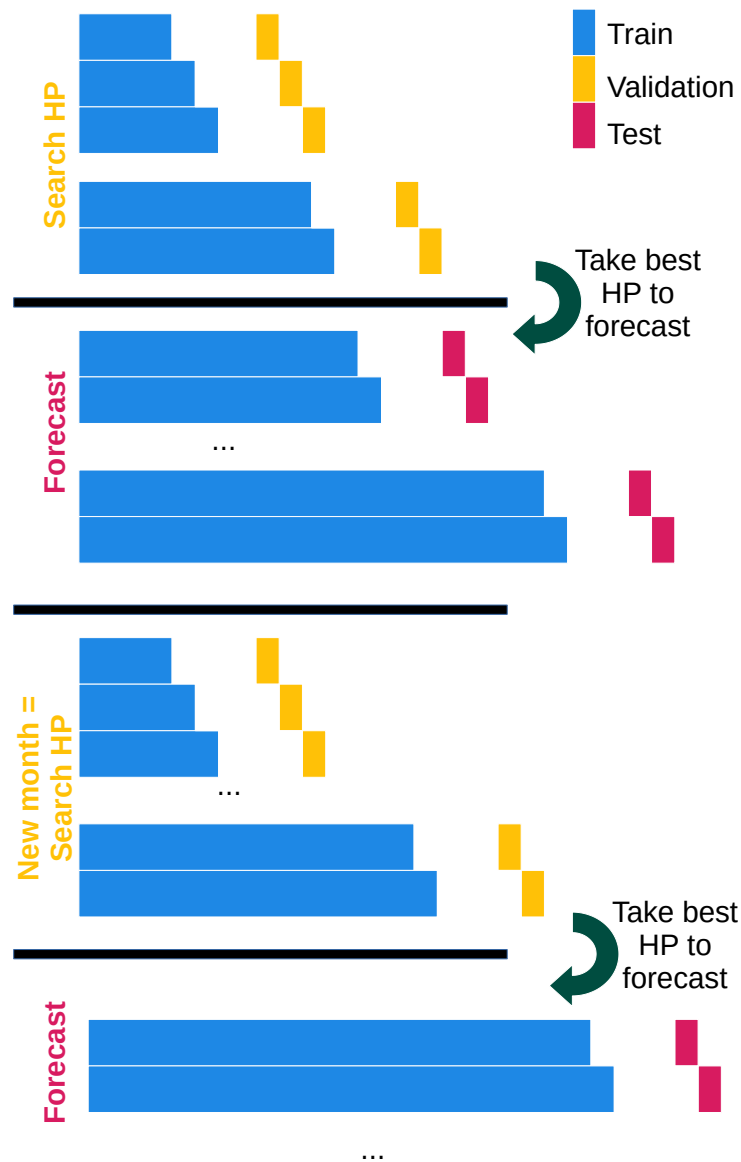


Fig. S1: Cross-validation scheme.

- XGBoost was chosen as a machine learning model. As a tree-based algorithm, it has been widely applied in time series forecasting, including SARS-CoV-2 forecasting.
- RC was selected as a neural network approach. Originally introduced by Jaeger et al. (19) and Maass et al. (20), RC is a distinct machine learning paradigm tailored for processing dynamic time series data (21). Unlike conventional recurrent neural networks (RNNs), RC consists of a randomly initialized recurrent layer (the reservoir), where only the output layer (read-out) is trained. This structure enables non-linear transformation of inputs into a high-dimensional space, capturing complex temporal dependencies while reducing computational cost compared to traditional RNNs (22). RC has been applied in various fields, including bird song analysis, language processing, epidemic modeling, power plant monitoring, internet traffic forecasting, and financial markets (24–28). However, selecting appropriate RC hyperparameters remains a challenge due to their impact on reservoir dynamics. To address this, we adopted the Genetic Algorithm (GA) optimization approach proposed by Ferté et al. (12).

Given the poor performance of heavier models such as Transformers in this context (12), they were not included in this study. The RC architecture used in this work featured 500 recurrent units, with both the input layer and reservoir connected to the output layer. Hyperparameter optimization was performed by evaluating the MAE of the median forecast across three RC instances with the same hyperparameter set. To assess the impact of hyperparameter optimization, we conducted an additional analysis using 20 RC instances instead of three.

For forecast evaluation, we computed the median forecast from 40 RC instances, each corresponding to one of the top 40 hyperparameter sets. Given the stochastic nature of RC due to randomly initialized reservoir connections, aggregating multiple forecasts helps improve stability and robustness.

10.1.4 Hyperparameter Optimization and Model Training

Hyperparameter selection was performed using random search for XGBoost and ENet and Genetic Algorithm (GA) for RC, as proposed by Ferté et al. (12). At model initialization, approximately 3,500 hyperparameter sets were evaluated, with 1,500 new evaluations added each month. Details on hyperparameter ranges are provided in Supplementary Table 4. The optimization process targeted:

- All hyperparameters of ENet
- Key hyperparameters of XGBoost, as identified by Yang et al. (68)
- Main hyperparameters of RC, based on the ReservoirPy documentation (69)

Table 4: Hyperparameter ranges

Model ¹	Hyperparameter	Range	Log ²
XGBoost	n_estimators	3;300	False
	max_depth	5;100	False
	learning_rate	1e-5;1	True
	subsample	0;1	False
	colsample_bytree	0;1	False
ENet	ridge	1e-10;1e5	True
	l1_ratio	0;1	False
RC	ridge	1e-10;1e5	True
	spectral_radius	1e-5;1e5	True
	leaking_rate	1e-5;1	True
	input_scaling	1e-5;1e5	False
	feature_selection ³	y, n	

¹ ENet : Penalized linear regression, RC : Reservoir computing

² For features with *Log = True*, features are log-transformed before being processed by GA or RS

³ One hyperparameter per feature

Following Ferté et al. (12), models were trained to forecast hospitalization variation rather than absolute hospitalization counts. However, all performance metrics and visual representations were reported on the raw hospitalization scale for interpretability.

To account for the non-stationary nature of COVID-19 time series data, we used a sliding window of one year for training. This approach was found to improve forecasting performance in previous research (12).

10.1.5 Feature selection

To explore the potential improvement in forecasting by incorporating expert knowledge, we developed a feature selection approach. Initially, all 409 predictors were included as input features. Feature selection was then performed using different methods depending on the model:

- For RC, feature selection was carried out via Genetic Algorithm (GA).
- For ENet, feature selection was embedded directly through L1 penalization.

- For XGBoost, feature selection occurs inherently as the algorithm selects relevant features during the tree-building process.

Additionally, we proposed selecting a subset of features a priori, based on expert knowledge. The chosen features were:

- Hospitalization data
- Number of positive RT-PCR tests
- Proportion of positive RT-PCR tests among individuals aged 60 and older
- IPTCC, a weather index specifically developed for SARS-CoV-2 forecasting
- Cumulative number of first-dose vaccinated individuals
- Number of people presenting to the emergency department with "COVID-19" mentioned in their Electronic Health Records (EHR)

Our hypothesis was that incorporating expert-driven feature selection would improve forecast accuracy by leveraging external information to better distinguish signal from noise. We expected this effect to be particularly pronounced in the early phase of the epidemic when limited data was available, and the signal-noise ratio is low.

10.2 Supplementary uncertainty quantification

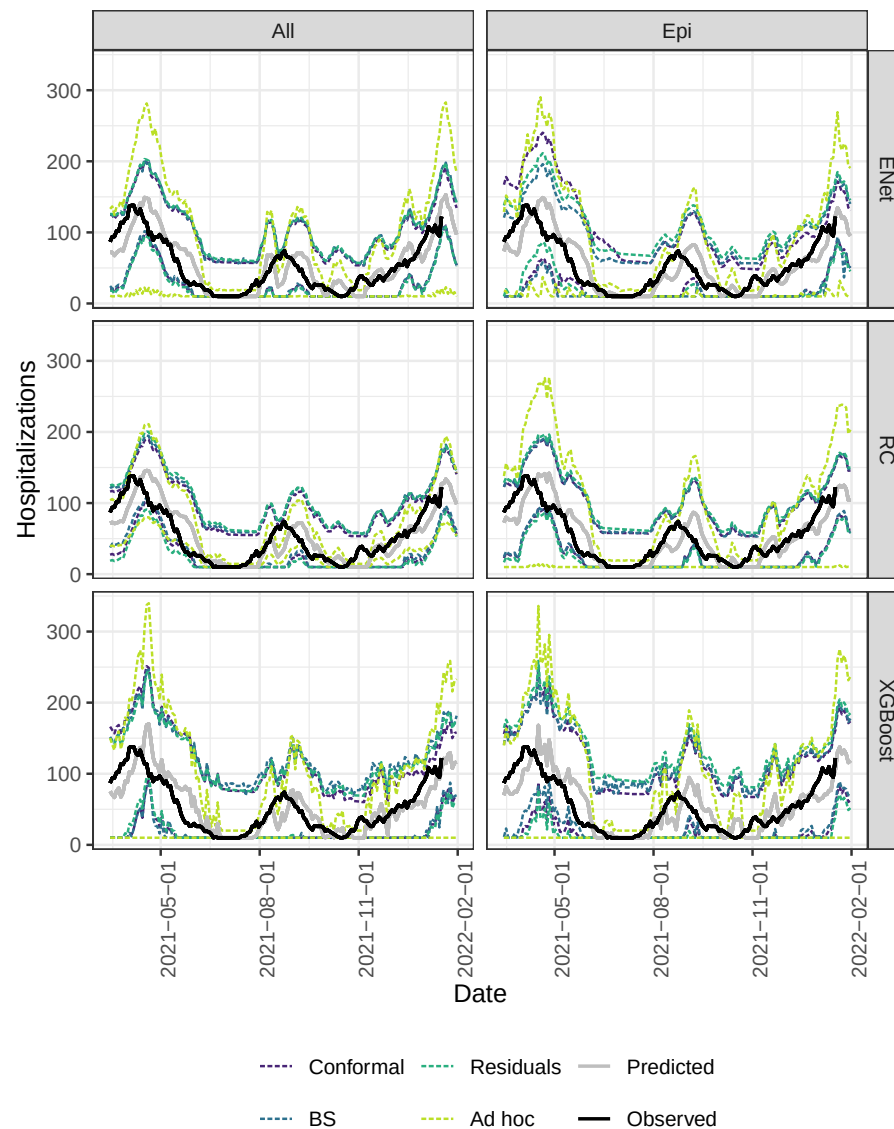


Fig. S2: Prediction intervals of Residuals, Bootstrap (BS), AgACI conformal inference (Conformal) and Ad hoc methods.