# Epidemiological forecasting using daily versus weekly aggregated data: Computational and practical implications

James Mba Azam[1,4*], Sam Abbott[1], Tumelo Sereo[2], Tobi Awodumila[3], Sebastian Funk[1], Carl Pearson[2,4]

**1** Center for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom,
**2** South African Center of Excellence for Epidemiological Modelling and Analysis, Stellenbosch University, Cape Town, South Africa,
**3** African Institute for Mathematical Sciences, Cape Town, South Africa,
**4** University of North Carolina, Chapel Hill, North Carolina, United States,

☯These authors contributed equally to this work.

¤Current Address: Dept/Program/Center, Institution Name, City, State, Country

†Deceased

¶Membership list can be found in the Acknowledgments section.

* james.azam@lshtm.ac.uk

## Abstract

Infectious disease forecasting is useful for public health decision-making, including resource allocation and the timing of outbreak response interventions. Forecast quality is impacted by the quality of the data inputs, including spatio-temporal resolution, but the extent is still an area of active research. Here, we evaluate differences between forecast performance and computational requirements between daily and weekly COVID-19 incidence data from the nine provinces of South Africa as forecasting inputs. We use the EpiNow2 R package as the core forecasting engine in an analysis pipeline that simulates in situ realtime forecasting with varying data streams. We evaluate generated forecasts against realized outcomes using the continuous ranked probability score (CRPS) and Effective Sample Size per second as measures of forecast performance and computational efficiency respectively. We find consistent trends in forecast performance (CRPS) and computational efficiency (ESS per second) across time, location, and forecast target resolution. Although both data inputs produced comparable forecast performance, the lower resolution weekly data inputs had lower computational efficiency. We describe a workflow for achieving comparable fits to data between daily and weekly data that can be applicable to similar forecasting problems. The extent to which these tradeoffs matter is context-specific, and varies depending on the public health question at hand. Therefore, the overall analytical costs, including collecting higher temporal resolution data and supplying greater computational resources, need to be weighed against the benefits of improved decision-making outcomes, like more prompt interventions and fewer false alarms. This analysis provides useful evidence towards establishing benchmark analyses for that kind of valuation of forecasting activities.

## Author summary

Author summary to be inserted

## Introduction

Infectious disease forecasts are increasingly used to inform public health decisions such as resource allocation and the timing of interventions during large outbreaks like influenza [1], Ebola [2, 3], and the COVID-19 pandemic [4]. Many approaches for making such forecasts exist, varying from highly mechanistic, capturing the biological processes, to statistical, relying on previously observed patterns in the data [5, 6]. However, the conditions under which a particular forecasting approach is most appropriate remain an open question [5].

Infectious disease forecasting models are often calibrated and confronted with data characterized by noise, incompleteness, and temporal delays in reporting and the interpretation of forecasts in this context is often of concern [4, 7, 8]. The temporal resolution of surveillance data is a key determinant of forecast performance with daily data being the gold standard as data aggregation lowers data fidelity and may mask short-term fluctuations [4]. Towards the latter parts of the COVID-19 pandemic, many health agencies shifted from daily to weekly case reporting to reduce the cost and burden of data collection [9]. Data is often aggregated for several reasons including removing week-day and weekend reporting biases.

Temporally aggregated data can be used for forecasting, but they also point to trade-offs between fidelity, bias reduction, and computational burden. Several recent studies have introduced algorithms to estimate transmission from aggregated incidence data. The EpiEstim R package uses an expectation-maximisation method for reconstructing daily case counts from weekly reports, allowing rapid estimation of time-varying reproduction numbers [8]. Another approach applied simulation-based methods using Approximate Bayesian Computation (ABC), recovering reproduction numbers accurately from weekly data, albeit with increased algorithmic complexity resulting from the repeated simulation of cases using ABC [10]. A Sequential Monte Carlo framework fit to renewal models either with day-of-week effects or with weekly aggregation has achieved good performance with lower root-mean-square error and CRPS than a chosen baseline model [11]. Although these methods demonstrate that aggregated data can be accommodated, their focus is primarily on retrospective inference (e.g., estimating reproduction numbers or elimination probabilities) rather than forecast performance on unseen data. Moreover, there is little discussion of how computational effort scales with data resolution or of how to systematically combine diagnostic checks with proper scoring rules to interpret performance.

The forecasting research community has established several methods for scoring [12, 13, 14] and evaluating [15] forecasts. The choice of score/metric depends on the forecast type (point, interval, and probabilistic) and ranges from traditional aggregate measures of distance to Proper Scoring Rules [5, 12, 13, 14, 16]. Forecast evaluation metrics such as the continuous ranked probability score (CRPS) quantify the distance between predicted and observed distributions and are widely adopted to assess forecast performance against observed data [12, 16, 17]. Recent work highlights that while existing metrics address theoretical concerns such as calibration and sharpness, they rarely address the operational challenges of generating forecasts, including reporting frequency and computational resourcing, pointing to a need for evaluation frameworks that bridge statistical evaluation and practical relevance [18].

Here, we benchmark a method for handling aggregated data in the EpiNow2 R package version 1.7.1 [19, 20]. Briefly, the EpiNow2 approach models daily latent infections using the renewal equation and maps them to observations using discrete convolutions of delay distributions, including the incubation period and reporting delays. We generate forecasts from daily and weekly COVID-19 reported cases in South Africa between March 2020 - August 2022 [21, 22], evaluate their accuracy using CRPS, assess convergence through rhat [23] and divergent transitions, and computational efficiency via effective sample size per second. By combining proper scoring rules with diagnostic measures, we quantify how forecast performance and computational requirements vary with data granularity and demonstrate how to achieve comparable fits across temporal resolutions. Our findings provide a transparent assessment of the trade-offs inherent in using aggregated data for real-time forecasting and offer practical guidance for balancing data collection efforts with forecasting needs, thereby advancing the development of evaluation standards and improving the utility of epidemic forecasts for public health decision-making.

# Methods

We developed a pipeline to get and clean observed data, perform the forecasts and save associated computational quantities such as MCMC convergence metrics and diagnostics and model run times, score the forecasts, and consolidate the outcomes for analysis and visualisation. In the following sections, we provide more detail on the individual parts of the pipeline.

## Data

We obtained and cleaned cumulative COVID-19 case counts for the nine South African provinces [21, 22]. The cleaning process focused on common reporting issues. First, two all-NA report dates (2020-03-27 and 2020-04-07) were removed up front. Next, the cumulative counts were converted to daily incidence per province by ordered differencing of non-missing values. We then corrected three reporting artefacts using a series of steps as follows:

- When a positive incidence case count had an equal neighbouring negative count (or vice versa), both values were set to zero.
- Using the cumulative series, we identified instances where negative daily incidence indicated that cumulative counts had been swapped; in these cases, we corrected the surrounding three-day window by converting the negative value to positive and adjusting the adjacent days so that the cumulative series remained strictly increasing.
- For any remaining negative daily values, we redistributed counts between a negative count and its positive neighbor by averaging the two values, ensuring both became non-negative while preserving the overall total incidence.

After cleaning the time series, we ensured each province had a non-negative daily time series for downstream aggregation and forecasting.

We aggregated the daily incidence time series into a weekly and rescaled weekly dataset. The weekly data were obtained by summing the daily counts in non-overlapping 7-day windows. The rescaled weekly data was a variant of the weekly data that represented each weekly total as if it occurred on a single day. This allowed us to evaluate models when all observations were spaced one week apart while preserving comparability with the other two datasets.

## Forecasting

We used EpiNow2's default model for estimating and forecasting reported cases (epinow) to generate 2-week ahead forecasts of COVID-19 reported cases for each of the nine provinces. We fit the model on 10-week chunks of the daily, weekly, and rescaled weekly data inputs, using 2-week sliding windows, and generated 2-week ahead forecasts for each slide. This approach allowed for consistent evaluation across different phases of the epidemic.

We configured the observation model in the default model to match the different observation patterns in the input data. For the daily data, the observation model accounted for day-of-week effects but this was turned off for the weekly and rescaled weekly data because the observations already reflect weekly totals.

EpiNow2 models the weekly aggregated data on a daily scale by accounting for the implicit missing dates and accumulating the modelled reported cases into weekly forecasts corresponding to the dates in the observed data. In this work, when we made forecasts from the weekly resolution data, we effectively set Saturday through Thursday as missing (NA's) and accumulated those reports to Fridays.

When fitting to the data in a sliding window, we allowed the model to be refitted until reasonable MCMC diagnostic and convergence criteria, defined below, were achieved. Stan (via cmdstanr with EpiNow2) was used for model fitting. We ran 4 chains in parallel, using 4 cores, with 5,000 posterior samples or 1500 iterations per chain. For each refit, we tuned stan's adapt-delta parameter to ensure efficient sampling and reliable convergence. Following recommendations from the Stan Community [24], the adapt-delta parameter was initiated at 0.80 and increased by 25% of the previous value for the next refit until it reached 0.99, which is the upper limit in stan. Based on the number of refits needed to go from adapt-delta of 0.8 to 0.99 in 25% increments, we determined that each slide could be refitted a maximum of 11 times. When the refit was reached without further improvements to the diagnostics, we returned the last model output. This approach allowed us to prevent the model from being refitted with minor improvements in the diagnostics.

We measured convergence using various diagnostic metrics. The bulk effective sample size (ESS) quantifies how many independent draws contain the same amount of information as the dependent sample obtained by the MCMC algorithm [23, 25, 26]. We limited bulk ESS to at least 100 per chain, hence a minimum of 400 for the 4 chains, based on recommendations with higher values of ESS indicating efficient mixing of the chains [23]. We additionally tracked rhat, which is a measure of MCMC convergence and limited it to 1.05 or less [23]. The number of divergent transitions during sampling, which indicates that the Hamiltonian Monte Carlo sampler has failed to accurately explore the posterior distribution due to irregular or complex geometry, potentially leading to biased inference was also limited to 25% of the 5000 samples [25, 26, 27]. We also tracked the model run times for computing efficiency metrics like the ESS per second. The diagnostic results were summarised per province and dataset, and time series of the ESS per second were plotted alongside reported cases to visualise sampling efficiency over time.

## Outputs

We saved outputs the forecast reported case counts per slide, MCMC diagnostics, including divergences, rhat, and bulk effective sample size per second, and model run times.

### Forecast scoring and evaluation

We evaluated forecast accuracy using the Continuous Ranked Probability Score (CRPS) given by

$$CRPS(F, x) = \int_{-\infty}^{\infty} (F(y) - 1\{y \geq x\})^2 dy$$

Where $F(\Theta)$ is the cumulative distribution function of $\Theta$, here, referring to the observed values, $x$, and forecast values, $y$, and 1 is the indicator function [12, 16, 28]. The CRPS is a proper scoring rule that quantifies the distance between the cumulative distribution of the forecast values and the observed case counts. CRPS was calculated on log-transformed forecasts and observations as this has been shown to produce more robust results [28]. The lower the CRPS value, the better. Forecasts were scored in four combinations within each slide:

- Daily forecast vs. daily observations – daily forecasts compared with daily reported cases.
- Daily forecast vs. weekly observations – daily forecasts aggregated to weekly totals and compared to weekly data.
- Weekly forecast vs. daily observations – weekly forecasts disaggregated on a daily scale and compared to the daily series.
- Rescaled weekly forecast vs. weekly observations – rescaled weekly forecasts compared to rescaled weekly counts.

We compared the default model's performance across the daily, weekly, and rescaled weekly data inputs using CRPS and ESS per second, computed as the effective sample size for a slide divided by the computation time in seconds. The two metrics represented forecast quality and computational efficiency respectively. Forecast scoring was facilitated by the scoringutils and scoringRules R packages [29, 30].

We plotted a time series of the CRPS and ESS/sec alongside the time series of log-transformed daily and weekly cases for easy interpretability. We also pooled together the results across time and computed the geometric mean relative to the model using daily data. To provide more details of model computational requirements, we also showed a time of the number of refits per slide per province and for all three input types.

### Data and code availability

The analysis used COVID-19 data from South Africa. The code for the analysis is available at https://github.com/epiforecasts/daily-vs-weekly-forecast-eval.

All analyses were done with R version 4.5.1 [31] and orchestrated with GNU make [32]. The SARS-CoV-2 incubation period was obtained from the epiparameter R package v0.4.1.900 [33], which stores a database of literature parameter estimates.

## Results

## Discussion

This study highlights the trade-offs of temporal aggregation in epidemiological forecasting. Daily data offer greater accuracy for short-term predictions, albeit at higher computational costs and diagnostic instability. Weekly data provide computational efficiency and smoother trends but sacrifice granularity and responsiveness to rapid

changes in case dynamics. Though we did not explicitly investigate this element, it stands to reason that the underlying surveillance system to support higher frequency reporting would also entail additional cost (e.g. more required storage space, more bandwidth to support more frequent access).

These results have two stakeholders. To the users of EpiNow2, it addressed questions of how EpiNow2 performs under degrading data resolution conditions and the computational requirements to achieve reasonable model fits. To the tool developers, these results inform ways to improve the underlying model to tackle the issue of degrading data conditions.

The findings emphasize the importance of aligning forecast resolution with the specific objectives of a public health response. For example, aggregated data may be suitable for long-term trend analysis, while high-resolution daily data are better for rapid outbreak detection and intervention planning.

The evaluation of scoring methodologies and diagnostics in this study also provides a framework for future analyses. Metrics such as CRPS, effective sample sizes, and divergence rates offer robust measures of forecast reliability and can be adapted to various modeling contexts. Though the daily data did generally provide better forecast performance, the algorithmic extensions to explicitly account for missingness kept the weekly resolution approach competitive. Such model enhancements may not always be practical, but this work suggests they can be valuable.

## Acknowledgements

The authors would like to acknowledge the International Clinics on Infectious Disease Dynamics and Data (ICI3D) faculty and particularly the Meaningful Modeling of Epidemiological Data (MMED) 2024 workshop faculty for their support during the incubation of this project.

## References

1. Doms C, Kramer SC, Shaman J. Assessing the Use of Influenza Forecasts and Epidemiological Modeling in Public Health Decision Making in the United States. Scientific Reports. 2018;8(1):12406. doi:10.1038/s41598-018-30378-w.
2. Meltzer MI. Modeling in Real Time During the Ebola Response. MMWR Supplements. 2016;65. doi:10.15585/mmwr.su6503a12.
3. Carias C, O'Hagan JJ, Gambhir M, Kahn EB, Swerdlow DL, Meltzer MI. Forecasting the 2014 West African Ebola Outbreak. Epidemiologic Reviews. 2019;41(1):34–50. doi:10.1093/epirev/mxz013.
4. Reich NG, Lessler J, Funk S, Viboud C, Vespignani A, Tibshirani RJ, et al. Collaborative Hubs: Making the Most of Predictive Epidemic Modeling. American Journal of Public Health. 2022;112(6):839–842. doi:10.2105/AJPH.2022.306831.
5. Lauer SA, Brown AC, Reich NG. Infectious Disease Forecasting for Public Health; 2020. Available from: http://arxiv.org/abs/2006.00073.
6. Banholzer N, Mellan T, Unwin HJT, Feuerriegel S, Mishra S, Bhatt S. A comparison of short-term probabilistic forecasts for the incidence of COVID-19 using mechanistic and statistical time series models; 2023. Available from: http://arxiv.org/abs/2305.00933.

7. Nash RK, Nouvellet P, Cori A. Real-time estimation of the epidemic reproduction number: Scoping review of the applications and challenges. PLOS Digital Health. 2022;1(6):e0000052. doi:10.1371/journal.pdig.0000052.

8. Nash RK, Bhatt S, Cori A, Nouvellet P. Estimating the epidemic reproduction number from temporally aggregated incidence data: A statistical modelling approach and software tool. PLOS Computational Biology. 2023;19(8):e1011439. doi:10.1371/journal.pcbi.1011439.

9. Conway E, Mueller I. Joint estimation of the effective reproduction number and daily incidence in the presence of aggregated and missing data; 2024. Available from: https://www.medrxiv.org/content/10.1101/2024.06.06.24308584v1.

10. Ogi-Gittins I, Steyn N, Polonsky J, Hart WS, Keita M, Ahuka-Mundeke S, et al. Simulation-based inference of the time-dependent reproduction number from temporally aggregated and under-reported disease incidence time series data. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2025;383(2293):20240412. doi:10.1098/rsta.2024.0412.

11. Steyn N, Parag KV, Thompson RN, Donnelly CA. A Primer on Inference and Prediction With Epidemic Renewal Models and Sequential Monte Carlo. Statistics in Medicine. 2025;44(18-19):e70204. doi:10.1002/sim.70204.

12. Gneiting T, , Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. 2007;102(477):359–378. doi:10.1198/016214506000001437.

13. Mitchell K, Ferro CAT. Proper scoring rules for interval probabilistic forecasts. Quarterly Journal of the Royal Meteorological Society. 2017;143(704):1597–1607. doi:10.1002/qj.3029.

14. Pic R, Dombry C, Naveau P, Taillardat M. Proper Scoring Rules for Multivariate Probabilistic Forecasts based on Aggregation and Transformation; 2024. Available from: http://arxiv.org/abs/2407.00650.

15. Tabataba FS, Chakraborty P, Ramakrishnan N, Venkatramanan S, Chen J, Lewis B, et al. A framework for evaluating epidemic forecasts. BMC Infectious Diseases. 2017;17(1):345. doi:10.1186/s12879-017-2365-1.

16. Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an interval format. PLOS Computational Biology. 2021;17(2):e1008618. doi:10.1371/journal.pcbi.1008618.

17. Funk S, Camacho A, Kucharski AJ, Lowe R, Eggo RM, Edmunds WJ. Assessing the performance of real-time epidemic forecasts: A case study of Ebola in the Western Area region of Sierra Leone, 2014-15. PLOS Computational Biology. 2019;15(2):e1006785. doi:10.1371/journal.pcbi.1006785.

18. Murphy AH. What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. Weather and Forecasting. 1993;8(2):281–293. doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

19. Abbott S, Hellewell J, Thompson RN, Sherratt K, Gibbs HP, Bosse NI, et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. Wellcome Open Research. 2020;5:112. doi:10.12688/wellcomeopenres.16006.2.

20. Abbott S, Hellewell J, Sherratt K, Gostic K, Hickson J, Badr HS, et al..

EpiNow2: Estimate Real-Time Case Counts and Time-Varying Epidemiological Parameters; 2025. Available from: https://zenodo.org/records/14899316.

21. Marivate V, Arbi R, Combrink H, de Waal A, Dryza H, Egersdorfer D, et al.. Coronavirus disease (COVID-19) case data - South Africa; 2020. Available from: https://zenodo.org/records/3819126.
22. Marivate V, Combrink HM. Use of Available Data To Inform The COVID-19 Outbreak in South Africa: A Case Study. Data Science Journal. 2020;19(1). doi:10.5334/dsj-2020-019.
23. Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner PC. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. Bayesian Analysis. 2021;16(2). doi:10.1214/20-BA1221.
24. Betancourt M. Identity Crisis; 2020. Available from: https://betanalpha.github.io/assets/case_studies/identifiability.html.
25. Stan Development Team. Stan Modeling Language Reference Manual, version 2.36. In: Stan Modeling Language Reference Manual, version 2.36; 2025.Available from: https://mc-stan.org/docs/2_36/reference-manual/index.html.
26. Betancourt M. A Conceptual Introduction to Hamiltonian Monte Carlo; 2018. Available from: http://arxiv.org/abs/1701.02434.
27. McElreath R. Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC; 2018.
28. Bosse NI, Abbott S, Cori A, van Leeuwen E, Bracher J, Funk S. Scoring epidemiological forecasts on transformed scales. PLOS Computational Biology. 2023;19(8):e1011393. doi:10.1371/journal.pcbi.1011393.
29. Bosse NI, Gruson H, Cori A, Leeuwen Ev, Funk S, Abbott S. Evaluating Forecasts with scoringutils in R; 2024. Available from: http://arxiv.org/abs/2205.07090.
30. Jordan A, Krüger F, Lerch S. Evaluating Probabilistic Forecasts with scoringRules. Journal of Statistical Software. 2019;90:1–37. doi:10.18637/jss.v090.i12.
31. R Core Team. R: A Language and Environment for Statistical Computing; 2025. Available from: https://www.R-project.org/.
32. GNU Make; 2006. Available from: https://www.gnu.org/software/make/.
33. Lambert JW, Kucharski A, Tamayo Cuartero C. epiparameter: Classes and Helper Functions for Working with Epidemiological Parameters; 2025. Available from: https://epiverse-trace.github.io/epiparameter/.