Evaluation of short-term multi-target respiratory forecasts over winter 2024-25 in England using sub-ensemble contribution analyses

Jack Kennedy[1], William Ferguson[1], Owen Jones[1], Steven Riley[1,2], Thomas Ward[1], Maria L. Tang[1], Jonathon Mellor[1]

1. UK Health Security Agency, London, UK
2. School of Public Health, Imperial College London, London, United Kingdom

*Corresponding author: jack.kennedy@ukhsa.gov.uk

Abstract

Background

Epidemic forecasting research often assesses ensembles and their component models using probabilistic scoring rules. Quantifying how individual models affect ensemble performance is challenging, particularly across multiple targets and spatial scales.

Methods

We present Winter 2024-25 forecasts of Influenza and COVID-19 hospital admissions in England and conduct a retrospective simulation using the operational component models. Forecasts were scored using the per capita weighted interval score (pcWIS) for counts and the ranked probability score (RPS) for ordinal trend direction. We compared operational retrospective forecasts, used generalised additive models (GAMs) to estimate the expected change in score from the inclusion of a model in a sub-ensemble, and used Pareto analysis to understand which sub-ensembles were Pareto-optimal across scoring rules.

Results

Nationally, the Influenza and COVID-19 operational ensembles achieved pcWIS of 5.20×10^{-7} and 3.98×10^{-7}, with RPS of 0.234 and 0.171 respectively. This corresponds to a 47% improvement in score versus sub-ensembles for Influenza pcWIS. However, Influenza operational ensembles were 22% worse than sub-ensembles, on average, when measured by RPS. For COVID-10, operational ensembles were 43% and 265% worse on average, than retrospective sub-ensembles by pcWIS and RPS, respectively. The sub-ensemble simulation showed individual models influenced the ensembles during different epidemic phases. The Pareto analysis

demonstrated that there can be a trade-off between relative direction and absolute count score optimisation.

Interpretation

Our analysis shows that UKHSA forecasts were well calibrated with observations and often had comparable performance to optimal ensembles. Our GAM and Pareto analyses inform model selection for future ensembles.

Author Summary

Forecasts of winter hospital pressures in England are an important tool for senior healthcare leaders. It is common practice to produce a forecasting ensemble, i.e. combine the predictions of multiple models to create a single, more accurate prediction. Forecasting teams should strive to produce the best forecast possible; one tool for this is retrospective evaluation over a forecasting season using proper scoring rules to assess performance. Our forecasts are constructed of two components, an epidemic trend direction estimate as well as forecast of hospital admission numbers. There are two main challenges we address. The first is understanding at which epidemic phase different ensemble contributions are most effective, the second is the joint optimisation of an ensemble for both trend direction and admission numbers forecast. We apply these methods to a variety of ensembles (sub-ensembles) based on our own modelling suite, and compare the sub-ensembles to our operational forecasts from the Winter 2024/25 season.

Introduction

Respiratory pathogens, such as SARS-CoV-2 and Influenza, place substantial pressure on health systems in winter through emergency care visits, hospital admissions, and bed occupancy. The UK Health Security Agency (UKHSA) is responsible for prevention, preparation, and response to infectious diseases in England. UKHSA provides the National Health Service (NHS) and other public sector organisations in England with short-term forecasts (14 day forecast horizon) of hospital admissions due to COVID-19 and Influenza. UKHSA also provide a range of other real-time modelling products across pathogens such as RSV and norovirus [1] each winter, as well as during outbreak response [2, 3].

There is a rich history of using a combination of models, referred to as an ensemble, to forecast Influenza [4], COVID-19 [5], and other respiratory disease indicators [6]. Model ensembling helps to address limitations of individual models, producing a consensus estimate which can be less biased and is more robust than contributing models. While ensembles generally outperform individual models [7], not all ensemble methods are equally performant [8, 9]. The short-term ensemble forecasts produced by UKHSA in

the winter 2023-24 season were shown to be useful, but with room for improvement [10]. In the past two winters, UKHSA has ensembled models via unweighted posterior stacking [11], which produces prediction samples from the ensemble. Many similar forecasting hubs and teams use quantile averaging techniques to construct ensembles; this is an accurate and practical approach [12] when stacking is impractical. For example, submitting a small collection of quantiles is computationally and operationally much simpler than submitting hundreds or thousands of posterior samples to a forecasting hub.

In real-time settings, the decision to include or exclude a model is typically guided by experience and judgement informed by performance and plausibility checks. This decision to include or not can be thought of as model weighting, where an excluded model within an ensemble is given zero weight. It is not straightforward to know in real time which models should be highly weighted at a given point in time [13, 14, 15]. However, knowledge of a model's past performance (from previous seasons or more recently) may help modellers to understand which models are most appropriate at different epidemic phases [16]. Further, individual model performance does not translate in a straightforward way to a change in ensemble performance [17]. We therefore aim to develop methods to understand which models are most useful in ensembles during different epidemic phases.

Over the winter 2024-25 season, UKHSA developed a suite of short-term forecast models for COVID-19, Influenza, RSV and norovirus. These forecasts were delivered to the health system, minsters, local government, and public health officials over winter to help inform planning and resource allocation. In this paper we present an evaluation of the prospectively reported COVID-19 and Influenza hospital admission forecasts used for decision making, as well as a retrospective exploration of individual models' contributions to sub-ensembles. By exploring sub-ensembles, we aim to understand how individual models contribute to forecast accuracy conditional on a range of possible ensembles, and how this contribution changes over time and geography to help inform future seasons ensemble composition.

Methods

Ethics statement

UKHSA have an exemption under regulation 3 of Sect. 251 of the National Health Service Act (2006) to allow identifiable patient information to be processed to diagnose, control, prevent, or recognise trends in, communicable diseases and other risks to public health.

Forecast targets

Our forecasting targets are key metrics for winter hospital pressures, identified by working with healthcare operational managers. In the Winter 2024-25 season, UKHSA forecast new test-positive hospitalisations for COVID-19, Influenza, respiratory syncytial virus (RSV), and norovirus, The basis of this paper will be to present and analyse forecasts of COVID-19 and Influenza hospital admissions, as the highest burden pathogens, with the most comprehensive ensembles.

Consistent and comprehensive data collection across hospitals in England is performed by the National Health Service. In secondary care settings, if an individual presents severe disease symptoms, a diagnostic test will be performed with results reported to UKHSA via the Second-Generation Surveillance System (SGSS). The data and forecasts are, for each date of a test, the number of unique inpatients with a new positive test taken within the past 24 hours. All forecasts are produced weekly, at daily granularity, with a 14-day forecast horizon. We also produce forecasts for RSV hospital admissions and Norovirus cases, but we do not discuss these in this paper.

The forecast targets discussed in this paper were produced at multiple spatial scales. Influenza and COVID-19 admissions were forecast at the level of integrated care boards (ICBs) and then aggregated to produce regional and national forecasts. An ICB is a sub-regional NHS body responsible for the management of local health services. ICBs are each nested within one of the seven NHS regions.

Our secondary forecast target approach is an epidemic trend direction assessment. We categorise an epidemic as one of “decreasing”, “stable”, or “increasing”, an ordinal classification. This is a simplification of our forecasts, but useful for many users. An upward change after the end of a 14-day period in the 7-day rolling average of the forecasting target by $\geq 20\%$ is considered an “increase”, while a downward change of $\geq 20\%$ is a “decrease”. Changes within these boundaries are considered “stable”. The relative frequency of each event across posterior samples gives us a trend direction probability.

Scoring rules and scoring analysis data

As our forecasts are probabilistic in nature, we generate entire probability distributions, via posterior sampling, to provide uncertainty quantification about future values of the target. This set of samples is converted into predictive quantiles; equi-tailed 50% and 90% intervals are reported to users, as well as the median forecast value. However, for evaluation purposes we store equi-tailed intervals with the following nominal coverages: 50%, 66%, 90% and 95%, as well as predictive medians. Proper scoring rules [18] are used to assess probabilistic forecasts in a way which encourages the

forecaster to provide an honest assessment of uncertainty about the value to be forecast.

The appropriate scoring rule to be used depends on the nature of the quantity of interest. For our predictions of admissions, we use the per capita weighted interval score (pcWIS). This is the weighted interval score (WIS) applied on a per capita scale, that is, y is replaced by $y' = y/(\text{population size})$ in the definition of WIS, with quantiles also divided by the population catchment size [19]. Scoring a transformed forecast can be more useful than scoring a raw forecast [20]. This per capita approach allows us to compare scores across different spatial geographies on the same scale, as some have larger population catchments than others e.g. nation versus region. The WIS is a weighted average of interval scores. If a $100(1-\alpha)\%$ prediction interval (l, u) is used to estimate y', then the per capita interval score is

$$pcIS_\alpha = (u-l) + \frac{2}{\alpha}(l-y')\mathbb{I}(y' < l) + \frac{2}{\alpha}(y'-u)\mathbb{I}(y' > u)$$

where $\mathbb{I}(\cdot)$ is the indicator function. If we have K prediction intervals as estimates of y', as well as a median estimate m, then the pcWIS is

$$pcWIS = \frac{2}{2K+1}\left(w_0|y' - m| + \sum_{k=1}^{K} w_k pcIS_{\alpha_k} \right)$$

with the usual choice of w_k being $\frac{\alpha_k}{2}$, and $w_0 = \frac{1}{2}$ [21].

The second score we use is for assessing the ordinal component ("increase", "stable", "decrease") of our forecasts using the ranked probability score (RPS) [22]. Let p_j be the forecast probability of category $j \in \{1, \dots, r\}$. Let the observed category be y with $o_j = 1(y = j)$. The ranked probability score is:

$$RPS = \frac{1}{r-1}\sum_{k=1}^{r-1}\left(\sum_{j=1}^{k} p_j - \sum_{j=1}^{k} o_j \right)^2$$

where both pcWIS and RPS are negatively oriented (lower values indicate better performance).

<u>Retrospective simulation experiment</u>

Due to operational constraints, not all the models used in the 2024-25 season were developed or in use at the start of the season, a common challenge for comprehensive

evaluation both for in-house modelling efforts and in collaborative hubs where submission is optional. In Figure 1 we see that some models were introduced in the late season, or models were used at different times. Models are sometimes dropped from an ensemble because either the predictions are unrealistic, or an issue which would have been too time consuming to fix before the deadline for forecasts to be shared with users. This means that scores aggregated from predictions over the season do not provide a fair comparison of model performance, as some models are present only in easier periods to forecast. Since only one ensemble is used per disease per week, we cannot be certain if we used the "best" ensemble, or if constituent models were appropriate.

To help answer these questions we conducted a large simulation experiment. We re-ran each of the models available by the end of the season, producing 14-day forecasts every week from the start of the season until the end of the season. Then, to understand how each model impacts the ensemble, we produced all possible sub-ensembles containing 3 models. Computing all possible ensembles would be highly computationally expensive: for a suite of k models there are $2^k - 1$ ensembles to be constructed at many spatial geographies. However, the method is general, and any ensemble size is theoretically manageable.

To further reduce computational expense, we used quantile averaging (mean quantile value or mean probability) to produce ensemble forecasts; a common practice for forecasting hubs [5]. In contrast, for operational forecasts during the season, we constructed an ensemble by stacking all posterior samples [10, 8]. We then score all models and ensembles via the *scoringutils* package [23].

<u>Generalised Additive Models (GAMs) and averaging to understand the effect of a model in an ensemble on score</u>

We smoothed the ensemble scores using methods available in the *mgcv* package, and used an averaging method (Equation 1) to estimate the effect of an individual model in an ensemble. This method is used to understand how a model influences the pcWIS as well as RPS. For all scores and diseases, our GAM structure is as follows:

$$\mathbb{E}\left(\log S_{r,l}(t)\right) \sim s(t) + s_r(t) + s_l(t) + \alpha_r + \beta_l$$

where $S_{r,l}(t)$ is the average score of ensemble r for a location level l (i.e. nation, region, ICB) over all forecast horizons for a given prediction start date t, $s(\cdot)$ is a global smooth for the score over time, $s_r(\cdot)$ and $s_l(\cdot)$ are ensemble and location level specific smooths – for each smooth we used penalised thin plate regression splines – and α_r and β_l are random effects for the ensemble and location level respectively. The structure reflects our forecasting unit: location and time. We use a Gaussian error structure after log-transforming our scores to prevent negative score predictions (pcWIS and RPS are non-negative, real values).

For a given forecasting unit, we define the effect $p_{m,l,t}$ of a component model m of ensemble r at time t and location level l over a fixed forecast horizon as

$$p_{m,l,t} = \frac{1}{|R(m)|} \sum_{r \in R(m)} S_{r,l}\,(t) - \frac{1}{|R|} \sum_{r \in R} S_{r,l}(t) \tag{1}$$

where R is the set of all considered ensembles – in our specific case this is all ensembles that can be made of size 3 – and $R(m)$ is the set of ensembles (of size 3) where m is a component model. This is inspired by a functional ANOVA [24] decomposition, and therefore could be extended to understand how combinations of models contribute to ensemble performance. This is, for a fixed location level, the average score from ensembles containing a given model, minus the average score across all ensembles. This is also conceptually similar to the use of Shapley values to understand component model importance [17].

The *gratia* package [25] allows us to generate quasi-posterior samples of $S_{r,l}(t)$ from the models fitted in *mgcv*; we therefore provide median estimates of the effect, as well as 90% credible intervals, by computing each component sum of Equation (1) on a sample-by-sample basis.

<u>Pareto front</u>

A unique aspect of our forecasting product is the two outputs: a 14-day forecast of target values, as well as the trend direction estimate. This can be viewed as a set of competing objectives, as a single ensemble may not optimise both scores. In addition, we have model outputs at different spatial scales. We therefore have 6 competing objectives: pcWIS at national, regional and ICB geographies, and RPS at those same geographies. Optimising performance for a single metric represents an oversimplification of practical forecasting challenges and diverse customer needs.

A Pareto Front (PF) [26] is a tool commonly used when there are multiple competing objectives and no single best solution. Members of a PF can be considered as candidates for a best solution to a multi-objective optimisation problem.

Consider our collection of ensembles R, and suppose $r, r' \in R$ are possible ensembles and $S_1(\cdot), S_2(\cdot), \ldots$ are negatively oriented scoring rules applied to the ensembles at a given forecasting unit. We say r dominates r' if $S_i(r) \leq S_i(r')\ \forall\ i$ and there exists j such that $S_j(r) < S_j(r')$. We then define the PF as

$$\mathrm{PF}(R) = \{r \in R\text{: there is no } r' \in R \text{ such that } r' \text{ dominates } r\}\ .$$

We perform two separate Pareto analyses. First, we perform a simplified Pareto analysis which weights the scores to reduce the dimensionality of the problem: we work with

$$\mathrm{RPS}^* = 0.5\mathrm{RPS}_{\mathrm{national}} + 0.25\mathrm{RPS}_{\mathrm{ICB}} + 0.25\mathrm{RPS}_{\mathrm{region}}$$

and we weight pcWIS to obtain pcWIS^* in the same way. This weighted approach is possible when the scores are on the same scale, which is inherent for RPS, and made possible by using pcWIS rather than WIS. This gives us a two-dimensional front. The weights are chosen to represent a belief that the national forecast is twice as important as the others, since this forecast is used by the most senior health officials in the country. However, the other forecasts are not unimportant, since they are used by a wide range of individuals. A "full" Pareto analysis which uses all scores at all spatial geographies is performed with results available in the supplementary materials (Table 3).

Results

Operational forecasts

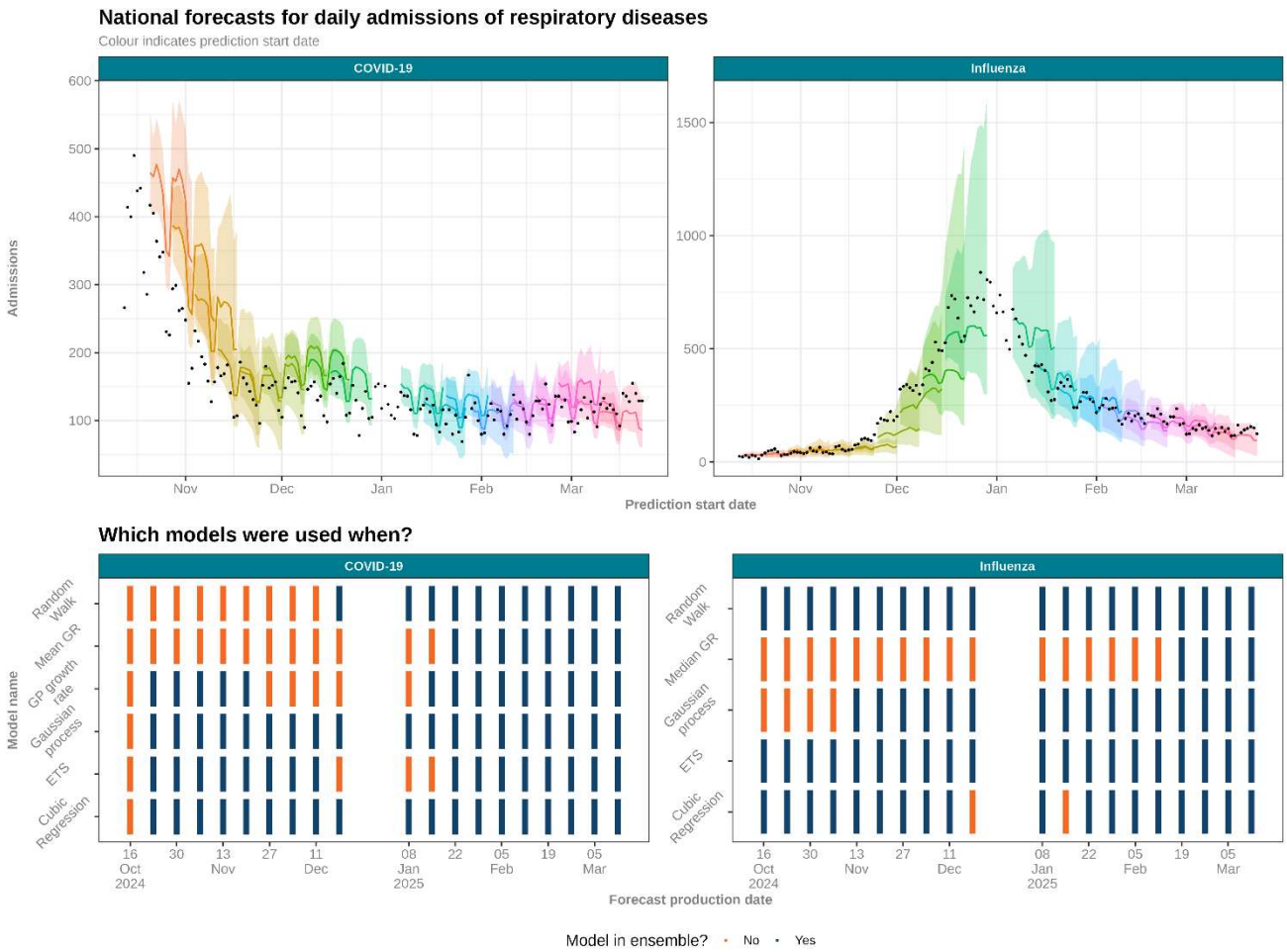

Figure 1. Top: UKHSA 14 day ahead predictions for COVID-19 and Influenza, with count of observed admissions shown. Forecasts are a median prediction (coloured lines) and a 90% prediction interval (colour-matched bands).

Bottom: Graphical representation of which models featured in the operational ensemble over time for each admission target.

Operational forecasts, and the component models used in each forecast, are presented in Figure 1. Our operational forecasts were able to accurately capture disease dynamics, with 14-day-ahead predictions generated on a weekly cadence. The gap in each set of forecasts is due to a break over the winter holiday period when no forecasts were produced. Forecasts presented to customers also included regional and ICB breakdowns. Figure D of the supplementary material shows time series of trend direction estimates.

Scoring analysis of operational forecasts

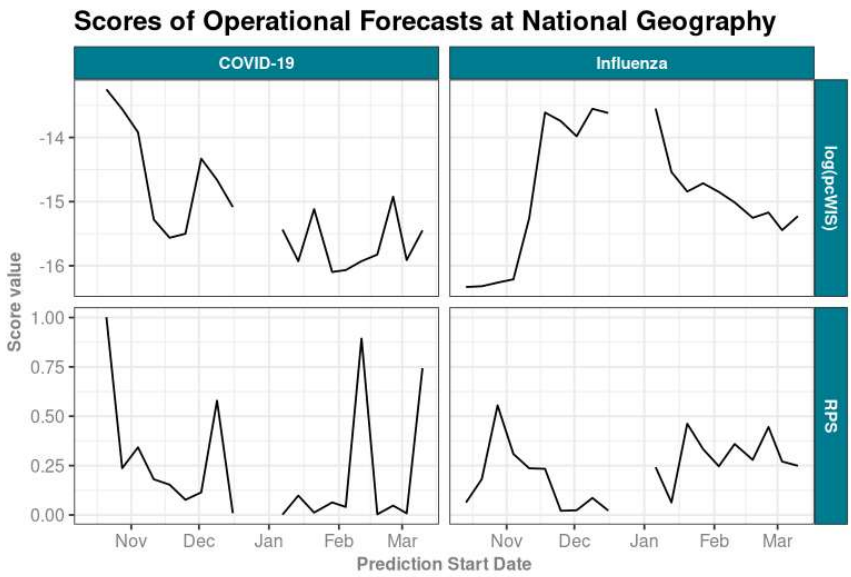

Figure 2. Average scores per prediction start date for COVID-19 and Influenza hospital admissions. Scores are averaged over forecast horizons.

For the real-time national forecast, we saw that for COVID-19 the pcWIS generally improved over the season (Figure 2). The slow-changing epidemic dynamics meant accurate forecasts were regularly produced [27]. For Influenza, the pcWIS increases up to the epidemic peak (late December), then decreases again. The number of hospital admissions around the peak of the epidemic was hardest to predict. Our forecast had large uncertainty around the epidemic peak. The predictive intervals were well-calibrated with the nominal coverage: 90% observed coverage statistics were close to the nominal coverage for all diseases and geographies (Supplementary Table 2). However, observed coverage at the 50% level was often higher than expected.

In terms of real-time trend direction, RPS was low for COVID-19, often near the theoretical lower bound of 0, with occasional spikes. This indicates the trend direction estimates were typically accurate, but the spikes indicate where trend direction estimation was relatively poor. Again, this is likely due to the simple, stable dynamics of

the epidemic. For Influenza, the models often performed well, obtaining small RPS values. The performance measure by RPS was consistent near the epidemic peak.

Retrospective simulation analysis of COVID-19 forecasts

Figure 3 shows the effects of individual models on ensemble performance for COVID-19 modelling. For pcWIS, we see that the GP growth rate model was detrimental to ensembles; the pcWIS increased by up to 50% around December, when hospital pressures approached their peak. For RPS, the effects of models on ensemble performance are smoother. We see that, again, GP growth rate is least useful around December-January. However, many other models become less useful as the season progresses. We see gradual upward trends in the performance of trend direction estimation for ETS, cubic regression and Gaussian process models. However, random walk and mean growth rate models show improving scores with time. The GP growth rate model is an experimental growth rate model, which can give unusual results when the growth rate is close to 0, which may explain poor performance, particularly around November to December. However, this did offer better-than-average performance towards the end of the season (see supplementary Figure A).

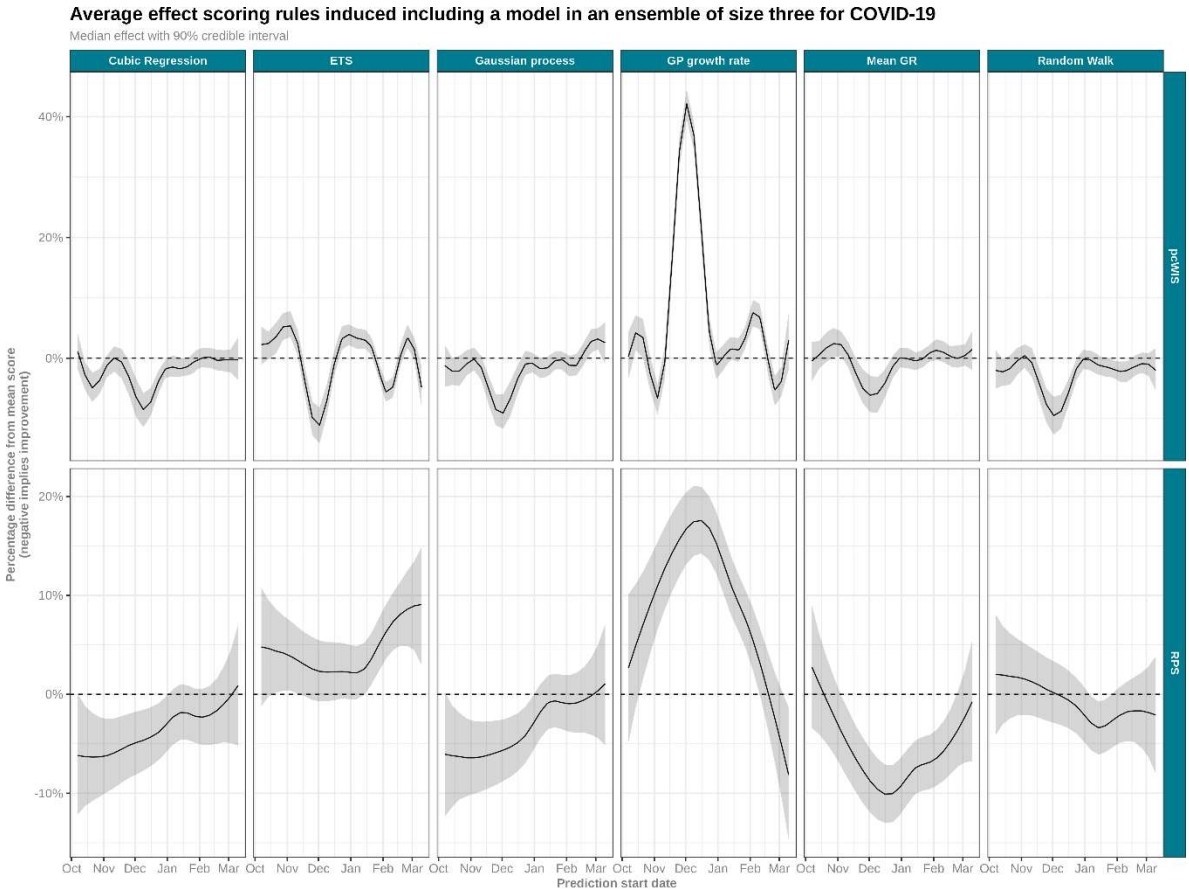

Figure 3. Estimated effect of including a model in an ensemble (expressed as a percentage change in pcWIS or RPS) for COVID-19 forecasts from the mean score at that point in time

Retrospective analysis of Influenza forecasts

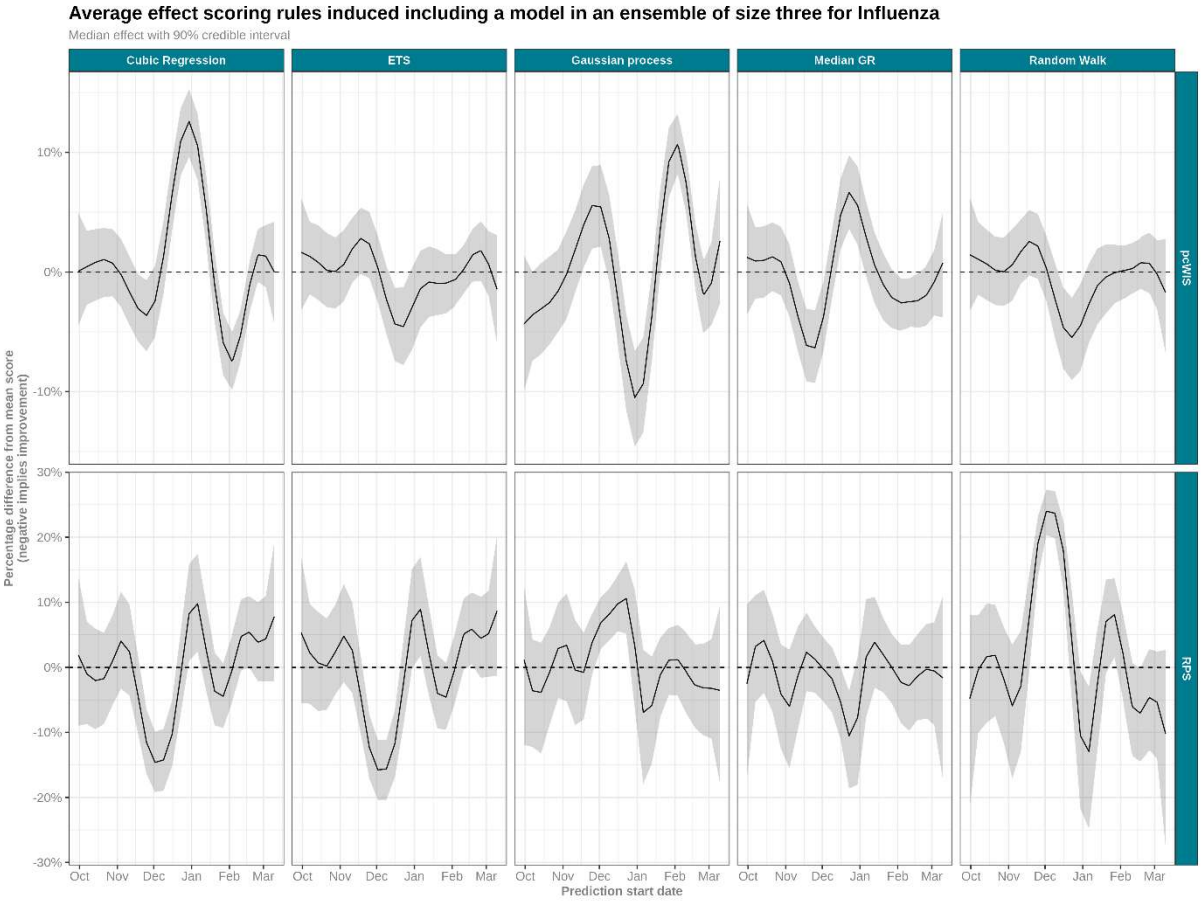

Figure 4. Estimated effect of including a model in an ensemble (expressed as a percentage change in pcWIS or RPS) for Influenza forecasts from the mean score at that point in time.

Figure 4 shows the effects of individual models on ensemble performance for Influenza modelling. We see that the improvement provided by models fluctuates over time. For example, we see that the cubic regression model is worst around the peak (late December – early January): cubic regression models tend to provide approximately exponential dynamics. However, ETS, random walk and Gaussian process models do well around the peak; all of these models have either "flat" predictions, or in the case of the Gaussian process, predictions which revert towards a constant mean. The median growth rate is an experimental model which uses growth rates from previous seasons to provide future forecasts. This provided reasonably good pcWIS contributions. In England, the peak in influenza admissions varies slightly from year to year [28], which may explain relatively poor performance around the peak for the median growth rate model.

Pareto Analyses

Weighted analysis

Since the weighted Pareto analysis is two-dimensional, we can plot the results. Note that in the Pareto analysis, we use an average score over the entire season. This allows us to understand which ensemble combination would have been best over the entire season, on average.

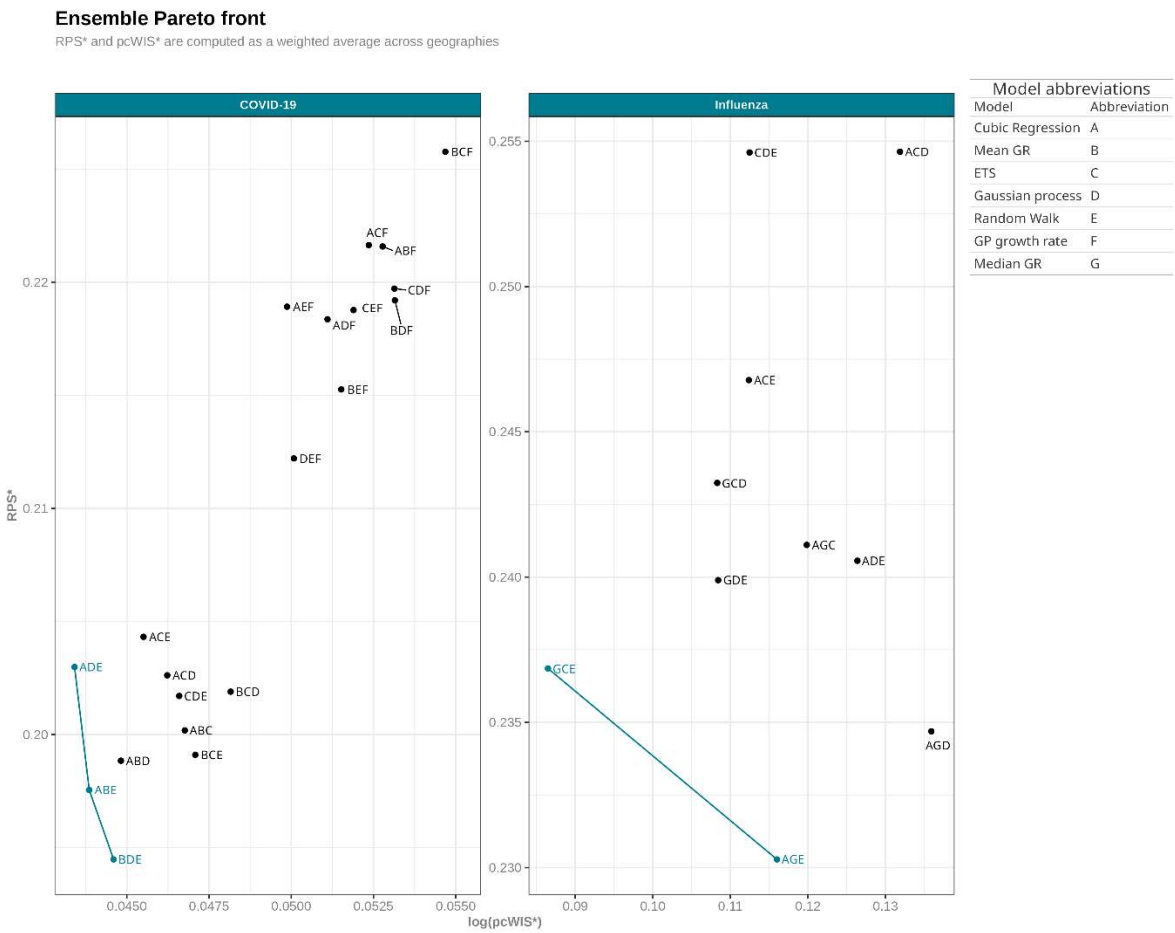

Figure 5. Pareto fronts for the weighted average scoring approach.

In Figure 5, we see that the PF for COVID-19 is constructed of three ensembles. Each ensemble contains a random walk model. Two contain one of {cubic regression, Gaussian process, mean growth rate}. Since the COVID-19 epidemic was near flat in the 2024-25 season, a random walk component in the ensemble is intuitive. The cubic regression, Gaussian process and mean growth rate models all behave quite differently. We also see a cluster of points far away from the PF. Every model in this ensemble contains a GP growth rate model, suggesting this model is detrimental to average ensemble performance. We see in the supplementary Figure A of retrospective forecasts that, on occasion, the GP growth rate forecasts exhibited unusual behaviour. However, the difference in score values is typically small between even the best and worst models, when averaging over the season.

The PF for Influenza under the weighted approach is small. The two members both contain random walk and median growth rate models. Both these models had 90% prediction intervals which almost always contained the true value (Supplementary Figure B), but which were perhaps too wide. The third model in each ensemble (cubic regression or ETS) often had quite narrow prediction intervals. The forecasts for both of these models around the peak could be quite poor but were reasonable in the run up to the peak and then after the peak.

<u>Comparison of retrospective forecasts to operational forecasts</u>

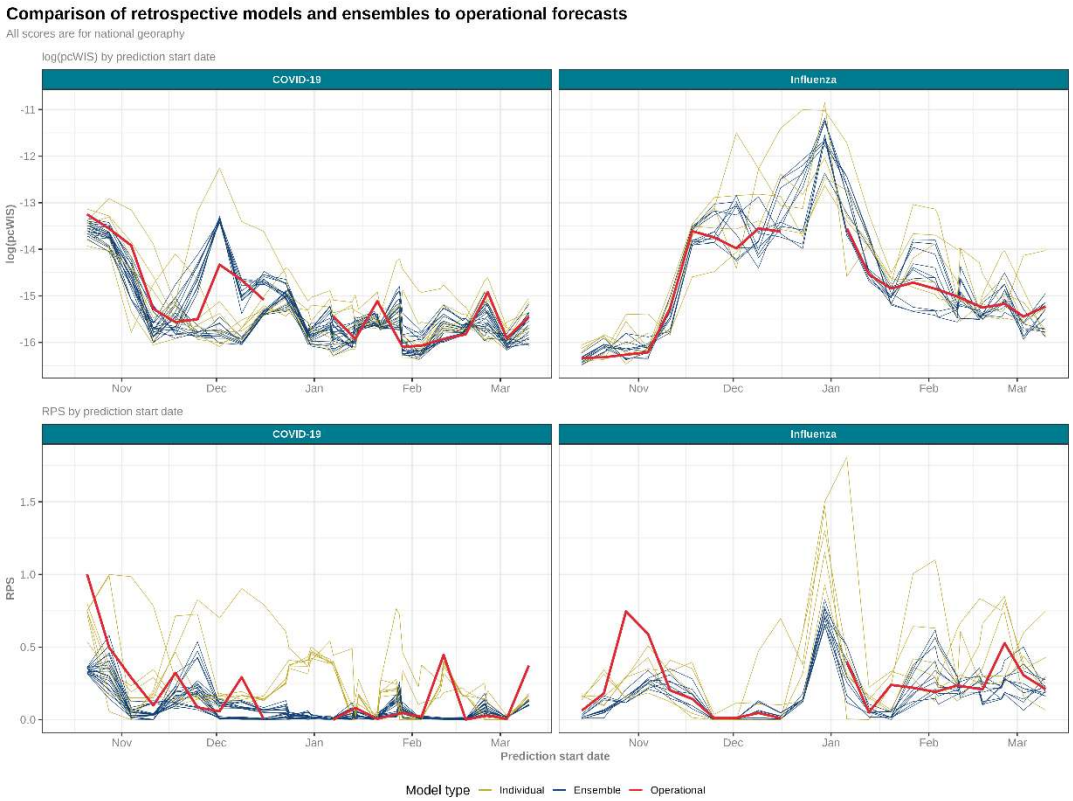

Figure 6. Comparison of how the operational ensemble performs relative to the retrospective 3-model ensembles and the retrospective individual models. The break in the red "Operational" line is due to the forecasting break which occurs over the Christmas and New Year.

Our operational ensemble could contain any number of models from 1 to the number of production models available at that time. Model inclusion in the ensemble is a subjective judgement, but we use a model in our weekly ensemble if the model predictions are epidemiologically plausible. We see in Figure 6 that, in general, our operational ensemble was of comparable performance to the retrospectively-fitted ensembles. The pcWIS values for the operational ensemble are usually around the average for retrospective ensembles; over the season the average pcWIS values for COVID-19 and Influenza operational forecasts were 3.98×10^{-7} and 5.20×10^{-7} respectively. For RPS, the mean RPS scores over the season were 0.242 and 0.236 for COVID-19 and Influenza respectively. Our operational ensembles occasionally gave large RPS values, especially for COVID-19. These spikes in RPS correspond to weeks

where the admissions were stable and we forecast a relatively low probability of a stable trend. That being said, our modal forecast was stable, so stakeholders received a correct forecast on the majority of weeks. Our Influenza RPS values for the operational ensemble were often marginally higher than most of the other retrospective ensembles. Minimum mean, overall mean, and maximum mean ensembles scores are available in Table 3 of the supplementary material.

Discussion

Many forecasts used to predict short-term disease dynamics are ensembles. Understanding the component models in such ensembles is important to build better forecasts and divert attention to maintenance and development of useful models. This paper provides a retrospective evaluation, tooling and methods to understand which models were useful at different epidemic stages and across an entire season for two pathogens.

Forecasts

Hospital admissions forecasts in the 2024-25 season were broadly accurate. The COVID-19 forecasts were especially accurate; however, the disease dynamics had limited complexity that season. Influenza forecasts were more challenging, especially when the peak was thought to be around the winter break in operations. The operationally-delivered ensembles are usually better than individual models, which is consistent with existing literature [5, 29].

There are numerous challenges to real time modelling including data quality issues and models failing due to bugs found in real-time. Code improvements to our modelling workflow over Summer 2024, most notably using the *targets* package, led to an easier-to-maintain collection of modelling pipelines compared to previous years [10]. Another year of experience producing forecasts under pressure led to better real-time decision making; combined with a more mature modelling suite and internal data quality monitoring tools, this resulted in all of our modelling outputs being delivered on time, or ahead of schedule without failure. We frequently delivered to customers at short notice due to the scheduling of critical system leader meetings.

Effects of models on ensemble scores

We have proposed a method for discriminating models in ensembles. Similar work includes LASOMO and LOMO methods to understand the importance of models in an ensemble [17]. A novel feature of our approach is to understand how a model may impact multiple different ensembles: the utilisation of GAM smooths can be used to understand the expected improvement in a scoring metric and provide uncertainty about these estimates. This approach gives a detailed understanding of which models

are most useful at which epidemic phases, relative to other candidate ensemble models. It is possible to estimate component model importance and use this to construct a weighted ensemble, which can have performance benefits [30]; this only shows us which models are useful rather than showing us models which have a negative effect on ensemble performance.

The use of a Pareto front to understand the trade-offs between different scoring rules for different quantities of interest is a contribution of this paper. This gives us a much broader view on the model choice problem than our GAM-based averaging approach alone.

Our approach is general and can be applied to any collection of forecasts generated at consistent intervals. The approach is sufficiently flexible to allow users to perform the analysis for different scoring rules, performance metrics or forecasting strata.

We presented a new GAM-based method for understanding how models influence ensemble performance, which can guide decisions about which models can or should be dropped from a modelling suite in favour of new modelling approaches. In particular, as a result of this analysis we have deprecated the COVID-19 Gaussian process growth rate model in favour of alternative approaches. Our Pareto analyses showed us which ensembles were useful when we have multiple scoring criteria which cannot be simultaneously optimised. This is a novel application of the Pareto front to retrospective evaluation of real-time modelling tools.

We also saw that our manually-chosen ensemble often performed worse than the retrospectively-fitted ensembles. However, observing the published forecasts, our predictions are reasonable. This is reassuring to users of UKHSA forecasting products that the predictions generated are high-quality. However, our operational ensembles often contained more models than the retrospectively-fitted ensembles (Figure 1). This shows that the largest possible ensemble, such as those used in forecasting competitions and hubs [31, 5, 32], are not a guaranteed route to improved predictive performance; a recent study advocates for ensembles with components providing complementary information, instead of simply aiming for as large an ensemble as possible [33]. Caution needs to be exercised when opportunistically including models in an ensemble: our ensembles should be diverse and have unique predictive features. This is seen in, for example, our influenza forecasts. It is well-documented in the literature that a diverse ensemble tends to lead to better results [8]. The ETS model tends to have a stable trend and narrow predictive bands, by contrast, the Historic GR model often had a non-stable trend with wide predictive intervals. Other influenza models had different behaviours to these models (supplementary Figure B).

<u>Limitations</u>

Limitations of our approaches are as follows. We have only used pure-statistical approaches in these ensembles. The ensemble could be improved with the addition of semi-mechanistic approaches which include parameters such as susceptible population sizes and infection hospitalisation rates. Semi-mechanistic approaches are an exciting addition to future ensembles [34, 35]. These can be complex to develop, but much of the complexity is waived by modern modelling software such as the *EpiNow2* package [36]. As with any scoring-based approaches, this work is retrospective, thus cannot be used in real time. However, scoring can be used mid-season to assess relative performance of models and contribution to model performance. Although there is evidence in the literature to suggest 4 or more models construct a robust ensemble [37], we chose ensembles of size 3 due to the limited number of models available for ensembling.

The (relatively) poor performance of our operational ensembles compared to retrospective ensembles can be attributed to a few factors. Firstly, not all models were available throughout the season, with some being introduced very late in the season when we had more time to experiment with methods. The 2024-25 season posed new challenges for the team as we employed new technologies to increase operational efficiency, but simultaneously were often under pressure to deliver forecasts ahead of regular cadence to meet decision makers' needs. This meant the team had less time to choose and experiment with different ensembles, as we had to provide forecasts to customers promptly. We were able to apply expert judgement to our forecasts and tune hyperparameters accordingly, and to include/drop models from our ensembles; however, this did not give us an advantage over many of our automatic runs this season (Figure 6).

Our future work will be to diversify our ensembles by including infection dynamic mechanism, as well as using models which use long-term historic data to use information such as seasonality to improve modelling. This will allow us to have a rich ensemble at the start of the season and provides a suite of modelling tools for use in the event of an incident or unexpected pandemic. We will expand RSV and norovirus ensembles to gain the benefits which ensembles can bring to performance. This will allow us to perform similar analyses for those diseases.

We have already removed models from our suite based on this analysis and could use this analysis to further improve future ensembles. Our ensembles are currently unweighted, thus one idea is to construct ensemble weights for a future season or dynamically within season [38, 39]. For example, the $p_{m,l,t}$ values could be used to calculate time-varying ensemble weights [8]; one challenge will be understanding how to use $p_{m,l,t}$ values for different scoring rules, forecasting outputs, or spatial resolutions to construct a single set of weights. As the size of our ensemble grows, we can use our

Pareto analyses as a first-pass for model evaluation: models which do not feature in a Pareto front should be considered for deprecation.

Acknowledgements

We thank UKHSA data operations colleagues for their work in the access, processing and maintenance of the data used in this paper. We would also like to thank Maximillian Ayling, Gregory Barnsley, Phoebe Asplin and Ian McFarlane of the Infectious Disease Modelling team for contributing to the operational delivery of forecasts during the Winter 2024-25 season.

Author contributions

Jack Kennedy: Conceptualisation; Methodology; Software; Formal analysis; Writing – Original Draft; Writing - Review & Editing; Visualization.
William Ferguson: Software; Writing - Review & Editing.
Owen Jones: Software; Writing - Review & Editing; Data curation.
Steven Riley: Writing – Review & Editing.
Thomas Ward: Writing – Review & Editing.
Maria Tang: Software; Writing - Review & Editing.
Jonathon Mellor: Conceptualisation; Methodology; Software; Validation; Writing - Original Draft; Writing - Review & Editing; Supervision.

Funding

This research did not receive any specific funding.

Data availability

An application for data access can be made to the UK Health Security Agency. UKHSA operates a robust governance process for applying to access protected data that considers:

- the benefits and risks of how the data will be used
- compliance with policy, regulatory and ethical obligations
- data minimisation
- how the confidentiality, integrity, and availability will be maintained
- retention, archival, and disposal requirements
- best practice for protecting data, including the application of 'privacy by design and by default', emerging privacy conserving technologies and contractual controls.

Access to protected data is always strictly controlled using legally binding data sharing contracts. UKHSA welcomes data applications from organisations looking to use protected data for public health purposes. To request an application pack or discuss a request for UKHSA data you would like to submit, contact DataAccess@ukhsa.gov.uk.

The modelling code for this project is available at https://github.com/jcken95/sub-ensemble-evaluation.

References

[1] J. Mellor, M. L. Tang, E. Finch, R. Christie, O. Polhill, C. E. Overton, A. Hoban, A. Douglas, S. R. Deeny and T. Ward, "An application of nowcasting methods: Cases of norovirus during the winter 2023/2024 in England," *PLOS Computational Biology,* vol. 21, no. 2, p. e1012849, 2025.

[2] M. L. Tang, I. S. McFarlane, C. E. Overton, E. Hani, V. Saliba, G. J. Hughes, P. Crook, T. Ward and J. Mellor, "Nowcasting cases and trends during the measles 2023/24 outbreak in England," *medRxiv,* 2025.

[3] C. E. Overton, S. Abbott, R. Christie, F. Cumming, J. Day, O. Jones, R. Paton, C. Turner and T. Ward, "Nowcasting the 2022 mpox outbreak in England," *PLoS Computational Biology,* vol. 19, no. 9, 2023.

[4] S. M. Mathis, A. E. Webber, T. M. León, E. L. Murray, M. Sun, L. A. White, L. C. Brooks, A. Green, A. J. Hu, R. Rosenfeld and others, "Evaluation of FluSight influenza forecasting in the 2021--22 and 2022--23 seasons with a new target laboratory-confirmed influenza hospitalizations," *Nature communications,* vol. 15, no. 1, p. 6289, 2024.

[5] K. Sherratt, H. Gruson, H. Johnson, R. Niehus, B. Prasse, F. Sandmann, J. Deuschel, D. Wolffram, S. Abbott, A. Ullrich and others, "Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations," *Elife,* vol. 12, p. e81916, 2023.

[6] N. Gozzi, C. Gioannini, P. Milano, I. Vismara, L. Rossi, M. Quaggiotto, S. Fiandrino, D. Paolotti, A. Vespignani, H. Johnson and others, "Performance evaluation of RespiCast ensemble forecasts for primary care syndromic indicators of viral respiratory disease in Europe during the 2023/24 winter season," *medRxiv,* 2025.

[7] V. Lopez, E. Cramer, R. Pagano, J. Drake, E. O'Dea, M. Adee and others, “Challenges of COVID-19 case forecasting in the US, 2020-2021,” *PLoS Computational Biology,* vol. 20, no. 5, p. e1011200, 2024.

[8] E. L. Ray and N. G. Reich, “Prediction of infectious disease epidemics via weighted density ensembles,” *PLoS Computational Biology,* vol. 14, no. 2, p. e1005910, 2018.

[9] N. Wattanachit, E. L. Ray, T. C. McAndrew and N. G. Reich, “Comparison of combination methods to create calibrated ensemble forecasts for seasonal influenza in the US,” *Statistics in Medicine,* vol. 42, no. 26, pp. 4696--4712, 2023.

[10] J. Mellor, M. L. Tang, O. Jones, T. Ward, S. Riley and S. R. Deeny, “Forecasting COVID-19, influenza, and RSV hospitalizations over winter 2023--4 in England,” *International Journal of Epidemiology,* vol. 53, no. 4, p. dyaf066, 2025.

[11] K. Sherratt, A. Srivastava, K. Ainslie, D. E. Singh, A. Cublier, M. C. Marinescu, J. Carretero, A. C. Garcia, N. Franco, L. Willem and others, “Characterising information gains and losses when collecting multiple epidemic model outputs,” *Epidemics,* vol. 47, p. 100765, 2024.

[12] K. S. Taylor and J. W. Taylor, “A comparison of aggregation methods for probabilistic forecasts of covid-19 mortality in the United States,” *arXiv preprint arXiv:2007.11103,* 2020.

[13] G. Claeskens, J. R. Magnus, A. L. Vasnev and W. Wang, “The forecast combination puzzle: A simple theoretical explanation,” *International Journal of Forecasting,* vol. 32, no. 3, pp. 754--762, 2016.

[14] E. L. Ray, L. C. Brooks, J. Bien, M. Biggerstaff, N. I. Bosse, J. Bracher, E. Y. Cramer, S. Funk, A. Gerding, M. A. Johansson and others, “Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States,” *International journal of Forecasting,* vol. 39, no. 3, pp. 1366--1383, 2023.

[15] A. V. R. Amaral, D. Wolffram, P. Moraga and J. Bracher, “Post-processing and weighted combination of infectious disease nowcasts,” *PLoS Computational Biology,* vol. 21, no. 3, p. e1012836, 2025.

[16] A. Adiga, G. Kaur, L. Wang, B. Hurt, P. Porebski, S. Venkatramanan, B. Lewis and M. V. Marathe, “Phase-Informed Bayesian Ensemble Models Improve Performance of COVID-19 Forecasts,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, Washington DC, USA, 2023.

[17] M. Kim, E. Ray and N. Reich, “Beyond forecast leaderboards: Measuring individual model importance based on contribution to ensemble accuracy,” *International Journal of Forecasting,* 2026.

[18] T. Gneiting and A. E. Raftery, “Strictly Proper Scoring Rules, Prediction, and Estimation,” *Journal of the American Statistical Association,* vol. 102, no. 447, pp. 359--378, 2007.

[19] S. Meakin and S. Funk, “Quantifying the impact of hospital catchment area definitions on hospital admissions forecasts: COVID-19 in England, September 2020-April 2021,” *BMC Medicine,* vol. 22, no. 1, p. 163, 2024.

[20] N. I. Bosse, H. Gruson, A. Cori, E. van Leeuwen, S. Funk and S. Abbott, “Scoring epidemiological forecasts on transformed scales,” *PLoS Computational Biology,* vol. 19, no. 8, p. e1011393, 2023.

[21] J. Bracher, E. L. Ray, T. Gneiting and N. G. Reich, “Evaluating epidemic forecasts in an interval format,” *PLoS Computational Biology,* vol. 17, no. 2, p. e1008618, 2021.

[22] E. S. Epstein, “A scoring system for probability forecasts of ranked categories,” *Journal of Applied Meteorology,* vol. 8, no. 6, pp. 985--987, 1969.

[23] N. I. Bosse, H. Gruson, A. Cori, E. van Leeuwen, S. Funk and S. Abbott, “Evaluating Forecasts with scoringutils in R,” *arXiv preprint arXiv:2205.07090,* 2022.

[24] B. Iooss and P. Lemaître, “A review on global sensitivity analysis methods,” in *Uncertainty management in simulation-optimization of complex systems: algorithms and applications*, Springer, 2015, pp. 101--122.

[25] G. L. Simpson, “gratia: An R package for exploring generalized additive models,” *Journal of Open Source Software,* p. 6962, 2024.

[26] I. Giagkiozis and P. J. Fleming, “Methods for multi-objective optimization: An analysis,” *Information Sciences,* vol. 293, pp. 338--350, 2015.

[27] L. A. White and T. M. León, “Forecastability of infectious disease time series: are some seasons and pathogens intrinsically more difficult to forecast?,” *medRxiv,* 2025.

[28] UK Government Official Statistics, “National flu and COVID-19 surveillance reports: 2025 to 2026 season,” 15 January 2026. [Online]. Available: https://www.gov.uk/government/statistics/national-flu-and-covid-19-surveillance-reports-2025-to-2026-season. [Accessed 21 January 2026].

[29] R. J. Oidtman, E. Omodei, M. U. Kraemer, C. A. Castañeda-Orjuela, E. Cruz-Rivera, S. Misnaza-Castrillón, M. P. Cifuentes, L. E. Rincon, V. Cañon and P. d. a. Alarcon, "Trade-offs between individual and ensemble forecasts of an emerging infectious disease," *Nature communications,* vol. 12, no. 1, p. 5379, 2021.

[30] G. Chowell, R. Luo, K. Sun, K. Roosa, A. Tariq and C. Viboud, "Real-time forecasting of epidemic trajectories using computational dynamic ensembles," *Epidemics,* vol. 30, p. 100379, 2020.

[31] S. Fiandrino, A. Bizzotto, G. Guzzetta, S. Merler, F. Baldo, E. Valdano, A. M. Urdiales, A. Bella, F. Celino, L. Zino and others, "Collaborative forecasting of influenza-like illness in Italy: The Influcast experience," *Epidemics,* vol. 50, p. 100819, 2025.

[32] A. Adiga, B. Hurt, G. Kaur, B. Lewis, M. Marathe, P. Porebski, S. Venkatramanan, A. Dukkipati, T. Gracious, S. Gupta and others, "A Multi-Team Multi-Model Collaborative Covid-19 Forecasting Hub for India," in *2023 Winter Simulation Conference (WSC)*, 2023.

[33] C. Dudley and M. Eisenberg, "Not All Accuracy Is Equal: Prioritizing Diversity in Infectious," 2025.

[34] S. Funk, A. Camacho, A. J. Kucharski, R. M. Eggo and W. J. Edmunds, "Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model," *Epidemics,* vol. 22, pp. 56--61, 2018.

[35] S. Bhatt, N. Ferguson, S. Flaxman, A. Gandy, S. Mishra and J. A. Scott, "Semi-mechanistic Bayesian modelling of COVID-19 with renewal processes," *Journal of the Royal Statistical Society Series A: Statistics in Society,* vol. 186, no. 4, pp. 604--615, 2023.

[36] S. Abbott, J. Hellewell, K. Sherratt, K. Gostic, J. Hickson, H. S. Badr, M. DeWitt, J. M. Azam, {EpiForecasts} and S. Funk, *EpiNow2: Estimate Real-Time Case Counts and Time-Varying Epidemiological Parameters,* 2025.

[37] S. J. Fox, M. Kim, L. A. Meyers, N. G. Reich and E. L. Ray, "Optimizing disease outbreak forecast ensembles," *Emerging Infectious Diseases,* vol. 30, no. 9, p. 1967, 2024.

[38] A. L. Buczak, B. Baugher, L. J. Moniz, T. Bagley, S. M. Babin and E. Guven, "Ensemble method for dengue prediction," *PloS one,* vol. 13, no. 1, p. e0189988, 2018.

[39] T. D. Do, T. D. Nguyen, V. C. Ta, D. T. Anh, T.-H. Tran Thi, D. Phan and S. T. Mai, "Dynamic weighted ensemble for diarrhoea incidence predictions," *Machine Learning,* vol. 113, no. 4, pp. 2129--2152, 2024.

Supporting information captions

Figure A. individual retrospective model predictions for COVID-19 hospital admission at national geography

Figure B. individual retrospective model predictions for Influenza hospital admissions at national geography

Figure C. Contributing variables to Pareto analysis.

Figure D: Trend direction probability estimates and [0, 1] normalised hospital admissions per disease

Table 1. Model details by disease.

Table 2. Comparison of nominal coverage of prediction intervals to observed coverage by disease and spatial granularity for operational ensembles. Statistics are an average for the entire season and where relevant, averaged across locations.

Table 3. Minimum, overall mean, and maximum values of retrospective ensemble scores averaged across the season; with operational ensemble average for comparison.

Table 4 Members of Pareto Front for full Pareto analysis (not using weighted averages) by disease.