ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Beyond forecast leaderboards: Measuring individual model importance based on contribution to ensemble accuracy

Minsu Kim *, Evan L. Ray, Nicholas G. Reich

School of Public Health and Health Sciences, University of Massachusetts, 715 North, Pleasant Street, Amherst, 01003, MA, United States of America

ARTICLE INFO

Article history:
Dataset link: https://doi.org/10.5281/zenodo.17954018, https://github.com/mkim425/replication_model-importance

Keywords:
Probabilistic forecasts
Ensemble
Model importance
Shapley value
COVID-19 forecasting

ABSTRACT

Ensemble forecasts often outperform forecasts from individual standalone models, and have been used to support decision-making and policy planning in various fields. As collaborative forecasting efforts to create effective ensembles grow, so does interest in understanding individual models' relative importance in the ensemble. To this end, we propose two practical methods that measure the difference between ensemble performance when a given model is or is not included in the ensemble: a leave-one-model-out algorithm and a leave-all-subsets-of-models-out algorithm, which is based on the Shapley value. We explore the relationship between these metrics, forecast accuracy, and the similarity of errors, both analytically and through simulations. We illustrate this measure of the value a component model adds to an ensemble in the presence of other models using US COVID-19 death probabilistic forecasts. This study offers valuable insight into individual models' unique features within an ensemble, which standard accuracy metrics alone cannot reveal.

© 2026 International Institute of Forecasters. Published by Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

1. Introduction

Forecasting is a crucial challenge across fields such as economics, finance, climate science, wind energy, and epidemiology. Accurate forecasts of future outcomes help individuals and organizations make informed decisions, enabling better preparedness and more effective responses to uncertainty. Ensembles (or combinations) of individual forecasts are considered the gold standard because they generally provide more reliable performance in terms of accuracy and robustness than most, if not all, individual forecasts (Clemen, 1989; Gneiting & Raftery, 2005; Lutz et al., 2019; Timmermann, 2006; Viboud et al., 2018).

In collaborative forecasting efforts, only the ensemble's forecasts are used or communicated. For instance, during the COVID-19 pandemic, the US COVID-19 Forecast Hub (https://covid19forecasthub.org/) combined probabilistic forecasts from over 90 research groups to produce ensemble forecasts that retain the structure of a predictive distribution for cases, hospitalizations, and deaths in the US (Cramer et al., 2022; Ray et al., 2023). These ensemble forecasts were used by the US Centers for Disease Control and Prevention (CDC) as official short-term forecasts to communicate with the general public and decision-makers (see Cramer, et al. (2022) and Centers for Disease Control and Prevention (2023)). The Intergovernmental Panel on Climate Change (IPCC) also uses a multi-model ensemble to assess robustness and uncertainty arising from differences in model structures and process variability (Lee et al., 2021) in its official reports, which policy-makers use as a foundation for climate-related decisions and strategies.

In this work, we develop a measure of the extent to which component models contribute to the ensemble's skill. This model importance metric is generally applicable with a range of measures of forecast skill, for example, the squared prediction error (SPE) for point predictions and

* Corresponding author.
E-mail address: minsu@umass.edu (M. Kim).

The numerical results presented in this manuscript were reproduced by the Editor-in-Chief on 22 December 2025.

https://doi.org/10.1016/j.ijforecast.2025.12.006
0169-2070/© 2026 International Institute of Forecasters. Published by Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Please cite this article as: M. Kim, E.L. Ray and N.G. Reich, Beyond forecast leaderboards: Measuring individual model importance based on contribution to ensemble accuracy. International Journal of Forecasting (2026), https://doi.org/10.1016/j.ijforecast.2025.12.006.

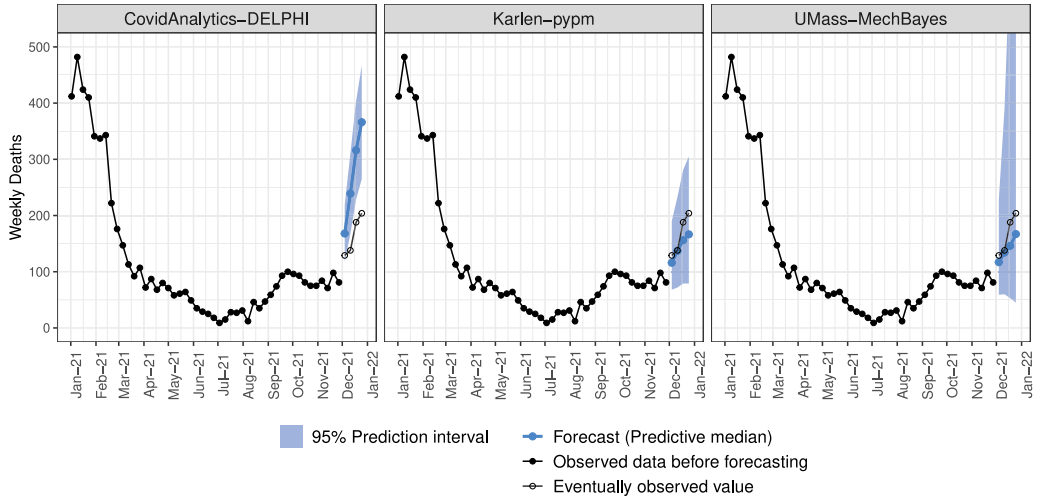


Fig. 1. Distributional forecasts of COVID-19 incident deaths at 1- through 4-week horizons in Massachusetts made on November 27, 2021, by three models. Solid black dots show historical data available as of November 28. Blue dots indicate predictive medians, and the shaded bands represent 95% prediction intervals. The open black circles represent observations that were not available when the forecast was made. The 95% prediction intervals of the UMass-MechBayes model (truncated here for better visibility of the observed data) extend up to 671 and 1110 for the 3-week and 4-week ahead horizons, respectively.

the weighted interval score (WIS) or log score for probabilistic forecasts (see details in Section 2.1). In the particular case of the expected SPE of a predictive mean, we show that our measure of model importance can be decomposed into terms measuring each component model's forecast skill as well as terms that can be interpreted as measuring how similar models' predictions are to one another. Through simulated examples, we demonstrate that the insights from this decomposition in the point prediction setting can be extended to other contexts, such as scoring probabilistic forecasts using WIS.

1.1. Motivating example

The predictions from different models may differ depending on the model's structure or the input data sources used. As an example, Fig. 1 shows the predictive distributions for incident deaths from three models submitted to the US COVID-19 Forecast Hub, along with eventually observed values. The quantile-based forecasts were made at 1-week through 4-week ahead horizons in Massachusetts on November 27, 2021. Here, two models under-predicted and one model over-predicted the outcomes. The CovidAnalytics-DELPHI model has narrow 95% prediction intervals. Still, its forecasts are more biased than those of the other two models across all four horizons, with especially large errors at forecast horizons of 3 and 4 weeks. On the other hand, the point estimates from the UMass-MechBayes model show less bias, but the predictive distributions are wide, especially for the 4-week-ahead incident deaths. The forecasts of the Karlen-pypm model are moderate in both bias and the width of the prediction interval. These different predictive abilities are reflected in evaluation scores used for probabilistic forecasts. At the 4-week forecast horizon, the Karlen-pypm model had the best WIS with 20.4, followed by UMass-MechBayes and CovidAnalytics-DELPHI with scores of 38.5 and 123.4, respectively. As discussed in more detail in Section 3.2, in this specific instance, the two models that were more accurate by traditional metrics actually proved to be less 'important' in the context of a multi-model ensemble because they were similar to multiple other submitted models. In particular, those two models, along with most models that contributed to the ensemble, had a small downward bias, which was partially offset by overpredictions from the CovidAnalytics-DELPHI model. The predictions that were too high, while less accurate than those from other models, were the only ones to over-predict the eventual outcome and therefore were important in offsetting a bias towards under-prediction across all the other models. This example motivates a closer examination of model importance within the ensemble forecasting framework.

1.2. Related literature

In the context of ensemble forecasting, some models will add more value than others. The problem of measuring each component model's impact on ensemble predictions bears methodological similarities to measuring variable importance in more standard regression-type modeling settings. Variable importance measures quantify the contribution of individual variables to the model's predictive performance. They are commonly used in model aggregation techniques such as Random Forest (Breiman, 2001) and Gradient Boosting (Friedman, 2001).

As a related concept, Shapley values from cooperative game theory measure, on average, how much each feature contributes to the predicted outcome across all possible coalitions of feature values. Building on the logic of the Shapley values, Giudici and Raffinetti (2021) proposed rank-based methodologies integrated with the Lorenz

Zonoid approach to quantify the contribution of individual predictors in a regression-like setting. Borup et al. (2024) also developed performance-based metrics tailored to the context of time series forecasting to achieve similar goals.

Beyond evaluating the contribution of predictors, the concept of Shapley values has been used to assess the value of component models in ensemble settings. For example, Pompigna and Rupi (2018) applied this concept to a method for weighting component models within an ensemble used to predict weekly/monthly fluctuations in average daily transport traffic. In the field of epidemic forecasting, Adiga et al. (2023) calculated Shapley-like values to determine the influence of each component model in the performance of a Bayesian Model Averaging ensemble at different forecasting phases of the COVID-19 pandemic.

While these prior studies primarily focus on point predictions, our framework is designed for probabilistic forecasts, incorporating uncertainty quantification to capture the full distributional behavior of predictive models, which we identify as a central contribution of our work.

The value of including diverse models in an ensemble has been discussed in numerous studies. Batchelor and Dua (1995) quantified diversity using the probability of a reduction in error variance, and applying this metric to data on US economic forecasts shows that ensembling is more beneficial when combining diverse forecasts than when combining similar ones. This idea has been supported by several follow-up studies, including Lamberson and Page (2012), Lichtendahl and Winkler (2020), Thomson et al. (2019) and Kang et al. (2022), which similarly emphasize that increasing the diversity of ensemble members improves overall ensemble performance. Furthermore, Brown et al. (2005) have demonstrated the benefits of both reducing individual forecast error and maximizing forecast diversity. The ambiguity decomposition demonstrated that this is a desirable feature of ensemble construction. This line of research provides a blueprint for improving ensemble accuracy by increasing the diversity of accurate individual forecasters. The model importance metric we introduce, while related to the ambiguity decomposition, provides a more direct interpretation in terms of ensemble forecast accuracy and can serve as a general probabilistic forecast metric for evaluating or ranking the contributions of each component model to an ensemble.

The remainder of this paper is organized as follows. In Section 2, we address some accuracy metrics commonly used for point forecasts and quantile forecasts and introduce our proposed model importance metrics, including two algorithms for model importance metric calculation in the context of probabilistic forecasting. We also discuss the decomposition of the model importance metric in the context of point forecasts. We follow this with simulation studies to demonstrate that the insights from the decomposition based on point forecasting generalize to the probabilistic forecast setting and to examine the effect of a component model's distributional forecast bias and dispersion on the importance metric in the leave-one-model-out algorithm setting. Section 3 provides results of applying the model importance metrics to real-world probabilistic forecast data from the US COVID-19 Forecast Hub. We present a case study investigating the relationship between the importance metric using the leave-one-model-out algorithm and WIS with quantile-based forecasts of incident deaths in Massachusetts in 2021. Subsequently, we compare all the metrics across a larger dataset. Section 4 discusses the limitations of our study and outlines potential directions for further investigations. Section 5 concludes the paper with a summary of the main findings and implications.

2. Methods

2.1. Accuracy metrics

Among the various forecast skill metrics developed to assess forecast quality, we focus on the mean squared prediction error for point predictions and the weighted interval score for probabilistic forecasts.

The squared prediction error (SPE) is defined as the square of the difference between the observed outcome y and the predicted value $\hat{y}_i$ from model i:

$$\mathrm{SPE}(\hat{y}_i, y) = (y - \hat{y}_i)^2 := e_i^2. \tag{1}$$

For the real-valued random variables Y and $\hat{Y}_i$, the expected squared prediction error (ESPE) is formulated as

$$\mathrm{ESPE}(\hat{Y}_i, Y) = \mathbb{E}[(Y - \hat{Y}_i)^2], \tag{2}$$

which quantifies the average squared discrepancy between the model's predicted values and the actual observed values. ESPE accounts for the general performance of the model by considering the average error across all possible predictions and penalizes larger errors more significantly than smaller ones by squaring the differences between predicted and actual values. A lower ESPE indicates a model that makes predictions closer to the actual values on average.

The weighted interval score (WIS) of a probabilistic forecast is an approximation of commonly used probabilistic scoring rules such as the continuous ranked probability score (CRPS) and pinball loss. The WIS is expressed in terms of predictive quantiles as follows (Ray et al., 2023):

$$\mathrm{WIS}(q_{1:K}, y) = \frac{1}{K} \sum_{k=1}^{K} 2\{\mathbf{1}_{(-\infty, q_k]}(y) - \tau_k\}(q_k - y), \tag{3}$$

where $q_{1:K}$ denotes a set of K distinct predictive quantiles, with q_k representing the kth quantile evaluated at the quantile level τ_k, for some positive integer K. For example, the predictive median corresponds to the quantile at $\tau_k = 0.5$. y denotes the observed value and $\mathbf{1}_{(-\infty, q_k]}(y)$ is an indicator function that equals 1 if $y \in (-\infty, q_k]$ and 0 otherwise.

The average of the accuracy metrics, either the SPE values or the WIS values, for model i across many modeling tasks represents the overall performance of model i.

2.2. Ensemble methods

Forecasts from multiple predictive models are aggregated to produce ensemble forecasts. The quantile-based forecasts are a common way to represent probabilistic forecasts, and the predictive quantiles generated from each component forecaster are used for the quantile-based ensemble forecast. Let different forecasters be indexed by i ($i = 1, 2, \ldots, n$) and let q_k^i denote the kth quantile from model i. The ensemble forecast value at each quantile level is calculated as a function of the component model quantiles

$$q_k^{\text{ens}} = f(q_k^1, \ldots, q_k^n). \tag{4}$$

Eq. (4) is also applicable to point forecasts, computing the ensemble prediction as a function of the point forecasts from component models. We note that the q_k^i used hereafter refers to the kth quantile of a model i for a specific forecasting task, which is a combination of the forecast location, date, and horizon.

We employed the mean ensemble method, where all component forecasters have equal weight at every quantile level.

2.3. Model importance metric

We propose two algorithms to evaluate a component model's contribution to an ensemble.

2.3.1. Leave all subsets of models out (LASOMO)

We use the Shapley value (Shapley, 1953), a concept from cooperative game theory.

Let N be the set of n players in a game, and v be a real-valued characteristic function of the game. The **Shapley value** ϕ_i of player i ($i = 1, 2, \ldots, n$) is defined as

$$\phi_i = \sum_{\{S:S\subset N,\ i\notin S\}} \frac{s!(n-s-1)!}{n!}\big[v(S\cup\{i\}) - v(S)\big], \tag{5}$$

where S is a coalition that consists of s players out of the total of n players, excluding player i ($s \in \{0, 1, 2, \ldots, n-1\}$). When $s = 0$, it indicates that $S = \emptyset$.

The characteristic function v is assumed to satisfy $v(\emptyset) = 0$ and for each subset S, $v(S)$ represents the gain that the coalition can achieve in the game. Accordingly $v(S\cup\{i\}) - v(S)$ in Eq. (5) represents the marginal contribution of player i to the coalition S, and its weight is computed by considering all possible permutations of players in S. An interpretation of this concept can be found in Section 1 of the supplement for further reference.

We calculate the importance metric of a component model in ensemble creation using the Shapley value. The n players and a coalition of s players in the game correspond respectively to the n individual forecasting models and a collection of s component models for an ensemble in our context. A proper scoring rule serves as the characteristic function. However, this choice of the characteristic function does not satisfy the assumption that the value of the empty set is zero, as any scoring metric cannot be applied to an empty set of forecasting models, which means no prediction. It is also not meaningful to assign a quantitative score to "no prediction". To avoid this difficulty, we modify Eq. (5) to eliminate the case of the empty subset. Consequently, the denominator in Eq. (5) is replaced with $(n-1)!(n-1)$. See also Section 1 of the supplement.

For a single forecast task τ, the importance metric (i.e., the contribution) of the component model i is calculated by

$$\phi_{i_\tau} = \sum_{\{S:S\subset N,\ i\notin S\}} \frac{s!(n-s-1)!}{(n-1)!(n-1)}\Big[\mu(F_\tau^{S\cup\{i\}}, y_\tau) - \mu(F_\tau^{S}, y_\tau)\Big], \tag{6}$$

where F_τ^A represents the ensemble forecast constructed based on the forecasts from models in the set A, y_τ denotes the actual observation, and μ represents a positively oriented proper scoring rule. The difference in μ reflects the extent to which the component models contribute to the accuracy of the ensemble. A positive value of ϕ_{i_τ} indicates that, on average across all coalitions of models, including the ith forecaster in the ensemble construction produces improved ensemble predictions. On the other hand, a negative value of ϕ_{i_τ} means that including the ith forecaster in ensemble construction degrades ensemble prediction performance on average.

The average of importance metrics of model i, ϕ_{i_τ}'s, across all tasks is the overall model importance metric of the model i:

$$\Phi(i) = \frac{1}{|\mathcal{T}|}\sum_{\tau\in\mathcal{T}} \phi_{i_\tau}, \tag{7}$$

where $\mathcal{T}$ represents a collection of all possible forecasting tasks, and $|\mathcal{T}|$ indicates its cardinality.

We note that the weight for a subset in Eq. (6) is calculated in the same manner as in the Shapley value formula in Eq. (5) that is the weighted average over all possible permutations of coalitions. In this formulation, the weight depends on the subset size. The rationale behind this approach is that as more models are involved, the marginal contribution of an additional model tends to decrease, since new models are more likely to provide redundant information in the ensemble. However, alternative weighting schemes are possible. For example, Adiga et al. (2023) used equal weights to all subsets regardless of their size, meaning that the importance metric is an evenly weighted average of the marginal contribution over the subsets.

2.3.2. Leave one model out (LOMO)

In addition to the Shapley value analysis using "all subsets", we measure the ensemble forecast performance when a single model is removed from the ensemble to see how much that component model contributes to improving the ensemble accuracy. Let S^{-i} denote the set of all models excluding model i, i.e., $S^{-i} = \{1, \ldots, n\}\setminus\{i\}$, where $i = 1, 2, \ldots, n$. Then, $F^{S^{-i}}$ represents the ensemble forecast built based on the forecasts from models in the set S^{-i}. That is, we remove the ith forecaster from the entire set of n individual forecasters and create an ensemble from the rest. Similarly, $F^{S^{-i}\cup\{i\}}$ represents the forecast from an ensemble model that includes all n individual forecasters. The importance metric of the component

forecaster i for a single task τ is measured by

$$\phi_{i\tau} = \mu(F_\tau^{S^{-i}\cup\{i\}}, y_\tau) - \mu(F_\tau^{S^{-i}}, y_\tau). \tag{8}$$

2.4. Decomposition of importance metric measured by the LOMO algorithm based on point predictions

In this section, we discuss components of importance metrics measured by the LOMO algorithm in the context of point predictions and their mean ensemble, which serve as a tractable starting point before extending to probabilistic forecasts.

We use the positively oriented squared prediction error (−SPE) to assess the accuracy of the predicted values, since −SPE facilitates a more intuitive interpretation of the resulting importance metric: a positive score indicates a beneficial impact on ensemble accuracy, while a negative score reflects a detrimental impact.

For n point forecasts $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ and the actual outcome y, the importance metric of model i is calculated by subtracting the negative SPE of the ensemble forecast made from the predictions of all models except model i from that of the ensemble forecast based on predictions from all n models, written as

$$\phi_i = -\left(y - \frac{1}{n}\sum_{j=1}^{n}\hat{y}_j\right)^2 + \left(y - \frac{1}{(n-1)}\sum_{j\neq i}\hat{y}_j\right)^2 \tag{9}$$

$$= -\frac{1}{n^2}e_i^2 - \frac{2}{n^2}\sum_{j\neq i} e_i e_j + \frac{2n-1}{[n(n-1)]^2}\left(\sum_{j\neq i} e_j^2 + 2\sum_{\substack{j\neq i\\ j<k}} e_j e_k\right), \tag{10}$$

where e_j indicates the prediction error between y and the predicted value $\hat{y}_j$ from model j ($j = 1, 2, \ldots, n$). Details of the process leading to Eq. (10) from Eq. (9) are available in the supplementary materials (see Supplemental Section 2).

The expected score is given as

$$\mathbb{E}(\phi_i) = -\frac{1}{n^2}\mathrm{ESPE}(\hat{Y}_i) + \frac{2n-1}{[n(n-1)]^2}\sum_{j\neq i}\mathrm{ESPE}(\hat{Y}_j) - \frac{2}{n^2}\sum_{j\neq i}\mathbb{E}(e_i e_j) + \frac{2(2n-1)}{[n(n-1)]^2}\sum_{\substack{j\neq i\\ j<k}}\mathbb{E}(e_j e_k). \tag{11}$$

The expected importance metric of model i consists of two kinds of terms. The ESPE terms capture the accuracy of the individual models. The first term shows that the expected importance of model i is lower when its ESPE is large, while the second term shows it is lower when the combined ESPE of the other models is small. The terms involving the product of prediction errors examine how two models' predictions relate to each other and to the actual observation. If the product of their errors is negative, it means those models' predictions are on opposite sides of the actual value (one overestimates while the other underestimates). The third term measures how much model i helps to correct the errors made by the other models; as the combined expected correction increases, the expected importance of model i increases. The last term indicates that when the forecast errors of different models are highly similar, model i is expected to be more important.

It is worth noting that while our decomposition is closely related to that from Brown et al. (2005) (see details in Section 2.1 of the supplement), our decomposition directly reveals that model i is rewarded if it is not correlated with others and if the other models are correlated with each other. We note that under the assumption of unbiased forecasts, the expected product of the errors corresponds to the covariance of the errors.

2.5. Simulation studies

In these simulation studies, we show that the decomposition insights developed in the point forecast setting remain applicable to probabilistic forecasts. We then explore the effect of bias and dispersion of a component model's predictive distribution on the importance of that model using the mean ensemble method in the LOMO algorithm setting. We focused on LOMO in these experiments because it closely aligns with the theoretical framework in Section 2.4 and is simple to interpret.

We created three simulation scenarios to assess model importance. The first two scenarios investigate model importance using component point forecasts and probabilistic forecasts with varying degrees of bias (Section 2.5.1). The third scenario investigates model importance with probabilistic forecasts with misspecified dispersion (Section 2.5.2). We assume that the truth values follow the standard normal distribution

$$Y_\tau \sim N(0, 1), \quad \text{for all } \tau \in \mathcal{T},$$

where $\mathcal{T} = \{1, \ldots, 1000\}$. For each of the probabilistic scenarios, we use 23 quantiles to represent the forecast distributions at the same quantile levels as in the data set used in the applications (see Section 3.1). We calculate the importance metric for each model based on individual observations, using the negative SPE for point forecasts and the negative WIS for quantile forecasts. As mentioned in Section 2.4, we adopt a positive orientation for a more straightforward interpretation, so that larger values reflect a more beneficial effect on the ensemble accuracy. The overall model importance is then taken as the average over $|\mathcal{T}| = 1000$ replicates, which approximates the expected importance metric.

2.5.1. Setting A: Relationship between a component forecaster's bias and importance

For the first scenario, we consider the following three-point forecasts:

$$\hat{y}_1 = -1, \quad \hat{y}_2 = -0.5, \quad \hat{y}_3 = b, \tag{12}$$

and in the second scenario, we assume that all three component forecasters produce normally distributed forecasts as follows:

$$F_{1,\tau} = N(-1, 1), \quad F_{2,\tau} = N(-0.5, 1), \quad F_{3,\tau} = N(b, 1), \tag{13}$$

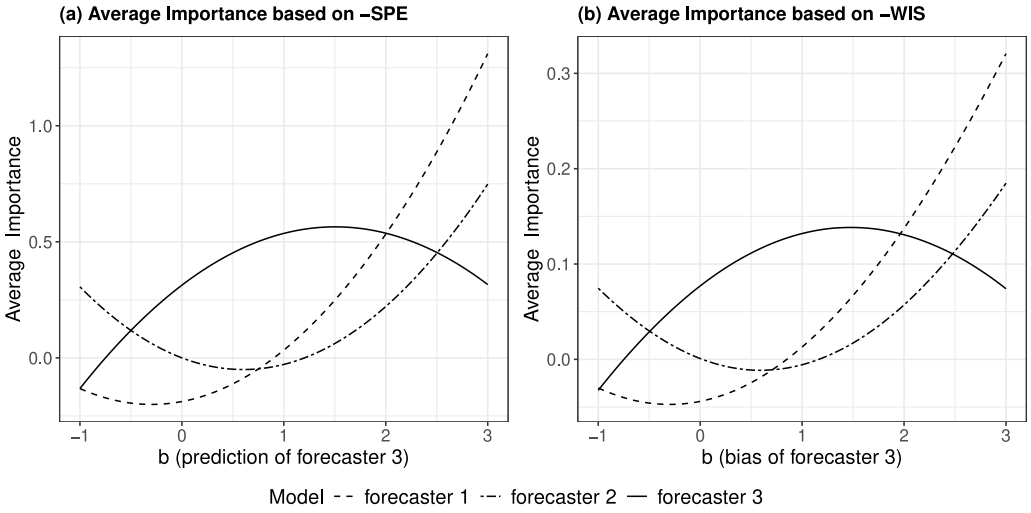


Fig. 2. Expected importance of three forecasters as a function of the prediction/bias of forecaster 3 in simulation settings: (a) $\hat{y}_1 = -1$, $\hat{y}_2 = -0.5$, and $\hat{y}_3 = b$ based on the negative SPE, (b) $F_{1,\tau} = N(-1, 1)$, $F_{2,\tau} = N(-0.5, 1)$, and $F_{3,\tau} = N(b, 1)$ based on the negative WIS, where $\tau = 1, \ldots, 1000$. The data generating process is $N(0, 1)$. The expected importance metrics were calculated and averaged over 1000 replicates of the forecasting experiments conducted at each value of b, incremented by 0.05 from -1 to 3.

where τ denotes the index of a generic replicate, with $\tau = 1, \ldots, 1000$. With the probabilistic component forecasts in Eq. (13), the ensemble forecast distribution is $F_\tau = N((b - 1.5)/3, 1)$. Note that the ensemble prediction is unbiased when $b = 1.5$. We changed the value of b from -1 to 3 in increments of 0.05 to observe how the importance of model 3 changes in both scenarios. Note that the point predictions correspond to the means of the probabilistic forecasters.

The simulation results show that the importance metric of each forecaster matches the calculation derived in Eq. (11) (Fig. 2(a)). Additionally, the general patterns of importance metrics observed for the three probabilistic forecasters closely align with the patterns seen with the point forecasters (Fig. 2(b)).

In both settings, the forecaster that produces the least biased forecast achieves the highest importance metric when all three forecasters have negative biases (i.e., biases in the same direction). However, when forecaster 3 has a small positive bias, unlike the other forecasters, it becomes the most valuable component model in the accurate ensemble creation, as it serves to correct the negative bias of the other component models. If forecaster 3 has a large bias ($b \geq 2$), then, although it is the only model biased in the opposite direction, forecaster 1 becomes the most important contributor to the ensemble. This is because forecaster 1 plays a more considerable role in offsetting that large bias compared to forecaster 2.

2.5.2. Setting B: Relationship between component forecaster dispersion and importance

In this simulation scenario, there are three probabilistic forecasts, each equal to a normal distribution with mean 0 and a different standard deviation:

$$F_{1,\tau} = N(0, 0.5^2), \quad F_{2,\tau} = N(0, 0.7^2), \quad F_{3,\tau} = N(0, s^2),$$

where τ denotes the index of a generic replicate, with $\tau = 1, \ldots, 1000$. In this setup, both forecasters 1 and 2 have predictive probability distributions that are underdispersed relative to the distribution of the data-generating process, which is $N(0, 1^2)$. With these component forecasts, the standard deviation of the ensemble forecast distribution is calculated as $(0.5 + 0.7 + s)/3$ (see details in Section 3 of the supplement). Note that the ensemble is correctly specified when $s = 1.8$. We changed s, the standard deviation of the forecast distribution produced by forecaster 3, from 0.1 to 3 in increments of 0.05.

Fig. 3 plots the expected or average importance metrics for the three forecasters as a function of the value of s. If the standard deviation of forecaster 3's predictive probability distribution is less than or equal to 0.5 (i.e., $s \leq 0.5$), then including that forecaster in the ensemble construction makes the ensemble's probabilistic forecast distribution narrower than not including that forecaster. This would make the ensemble's prediction very different from the truth, resulting in the forecaster having the lowest importance metric among all models. Starting from $s = 0.7$, forecaster 3 becomes the most important model, as the standard deviation of the ensemble's forecast distribution with that model included approaches that of the truth distribution more closely than the ensemble without it, as s increases. For $s \geq 1$, the predictions of F_3 become more and more overdispersed as s grows, and this large variance brings the dispersion of the ensemble close to the truth; however, beyond a certain point, the ensemble predictions become more dispersed than the truth. Thus, forecaster 3 maintains its top ranking in importance until s reaches approximately 2.4. Thereafter, the ensemble formed without forecaster 1 shows high forecast dispersion, leading forecaster 1 to have the highest importance metric among all models.

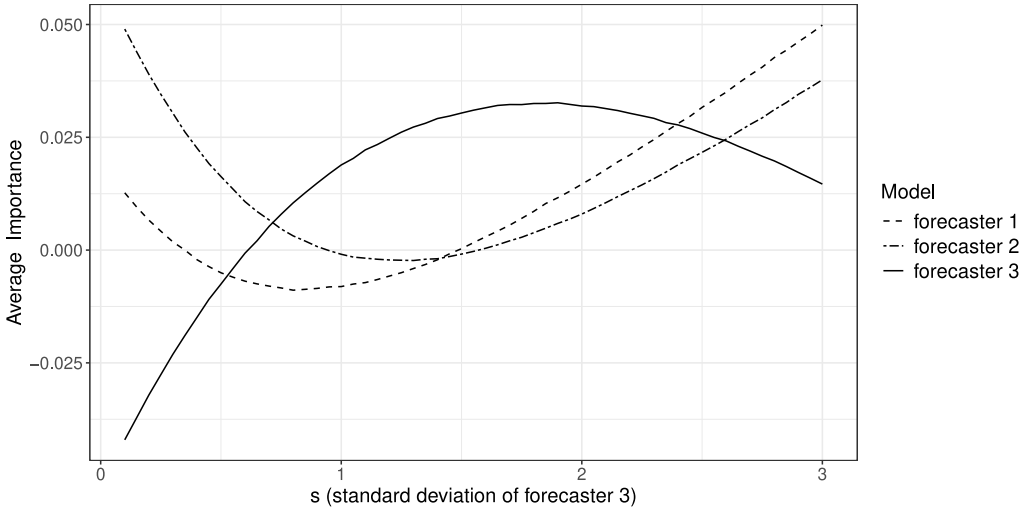


Fig. 3. Expected importance of three forecasters as a function of dispersion of forecaster 3 in the simulation setting: $F_{1,\tau} = N(0, 0.5^2)$, $F_{2,\tau} = N(0, 0.7^2)$, and $F_{3,\tau} = N(0, s^2)$ based on the negative WIS, where $\tau = 1, \ldots, 1000$. The data generating process is $N(0, 1)$. The expected importance metrics were calculated and averaged over 1000 replicates of the forecasting experiments conducted at each value of s, incremented by 0.05 from 0.1 to 3.

3. Application

In this application, we used probabilistic forecasts of COVID-19 deaths in the United States to evaluate each component model's contribution to probabilistic ensemble forecasts produced by the mean ensemble method. In a case study (Section 3.2), we used the LOMO measure to provide clearer illustrations of the key intuitive insights, as LOMO offers more straightforward interpretability than LASOMO. A more extensive application is presented in Section 3.3, where both LASOMO and LOMO algorithms were applied and compared.

The code used for loading data and conducting all analyses and simulations is archived on Zenodo[1] for reproducibility. The latest version of the associated code and data are available on GitHub.[2]

3.1. Data

The forecast data employed in this analysis were obtained from the US COVID-19 Forecast Hub, which collected short-term quantile forecasts of COVID-19 deaths from various models developed by academic, industry, and independent research groups, from its launch in April 2020 (Cramer et al., 2022) through April 2024. The submitted forecasts were provided using 23 quantiles (0.01, 0.025, 0.05, 0.10, 0.15, …, 0.90, 0.95, 0.975, 0.99). The death data on COVID-19 from Johns Hopkins University Center for Systems Science and Engineering (JHU-CSSE) were used as the ground truth data (Dong et al., 2020).

3.2. Case study: Relationship between importance metric and WIS with data for deaths in massachusetts in 2021

Our first analysis is a small case study designed to investigate the relationship between model importance calculated with the leave-one-model-out (LOMO) algorithm and model accuracy measured by the negative WIS. The forecasts analyzed were a subset of all forecasts from the Forecast Hub, including only 4-week-ahead forecasts of new deaths in Massachusetts for every week in 2021. The only models included were those that had made real-time forecasts for every week in 2021, to avoid complications arising from missing forecasts. We also excluded models that were ensembles of other models in our pool. This led to a set of 9 individual models. In building ensemble models, an equally weighted mean was used at each quantile level.

In Massachusetts in 2021, the importance metrics of component models were correlated with model accuracy as measured by −WIS. Specifically, the more accurate a model's predictions were on average (as the value of negative WIS increases, it indicates higher accuracy), the higher the importance that model had (larger values indicate more important forecasts) (Fig. 4). However, in certain weeks, there are still models with high importance metrics despite low accuracy (i.e., low negative WIS), suggesting that other factors determine a component model's importance. An example is the forecasts with a target end date of December 25, 2021, where the CovidAnalytics-Delphi model was the most important contributor to the ensemble. Still, as measured by negative WIS, it was also the least accurate for that forecast task (Fig. 5(a)). This is because, while this model had a large positive bias, it was the only one to show a bias in that direction on this

[1] https://doi.org/10.5281/zenodo.17954018
[2] https://github.com/mkim425/replication_model-importance

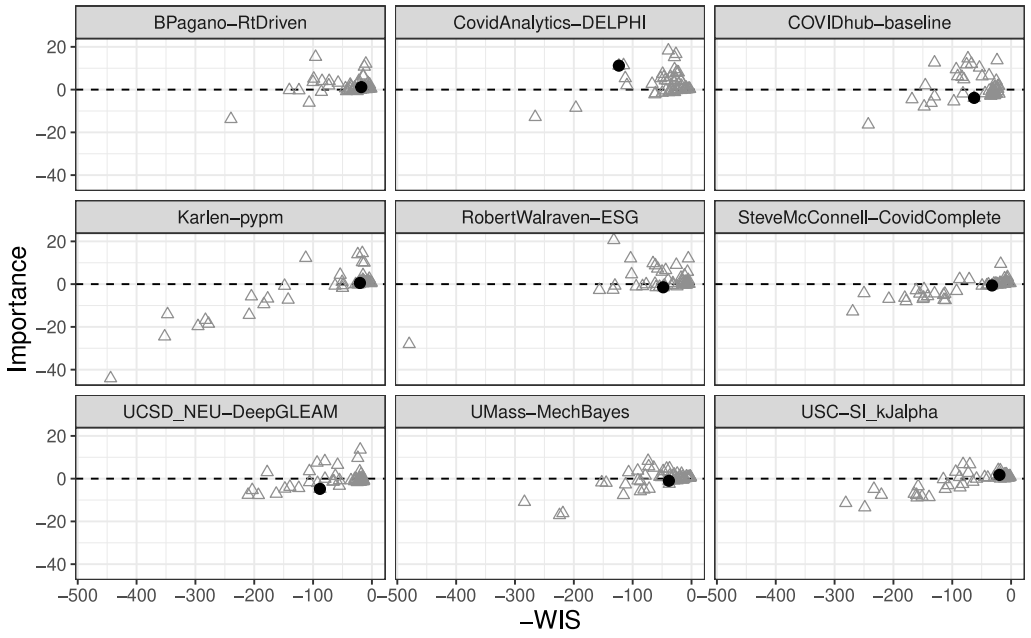


Fig. 4. Model importance versus negative WIS by model for all weeks in 2021. Each triangle represents a pair of negative WIS (*x*-axis; larger values indicate more accurate forecasts) and an importance metric (*y*-axis; larger values indicate more important forecasts) for a week in 2021. Solid black circles represent negative WIS and importance metric pairs evaluated for the one week ending December 25, 2021 (see more details in Fig. 5). The horizontal dashed lines indicate the value of zero. The importance of an individual model as an ensemble member tends to be positively correlated with the value of negative WIS; that is, the importance metric is positively correlated with the model's prediction accuracy.

particular week (Fig. 5(b)). That bias made this model an important counterweight to the other models, and adding it to an ensemble shifts the predictive median towards the observed data. This illustrates that component forecasts that are not accurate relative to other forecasts but offer a unique perspective can still play an important role in an ensemble.

3.3. Importance metrics measured by different algorithms

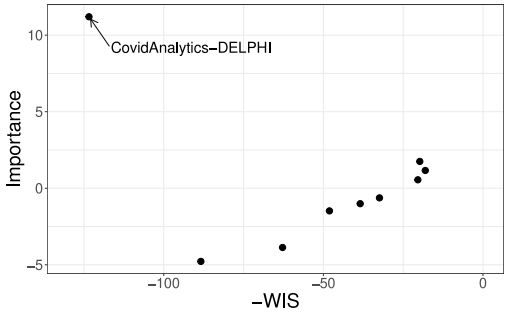
For this application, out of a total of 56 non-ensemble individual models that submitted forecasts of COVID-19 deaths to the Forecast Hub, we chose 10 models that submitted over 90% of the possible individual predictions for deaths across 50 states in the US and 1 through 4 horizons for 109 weeks from November 2020 to November 2022 (Table 1).

As mentioned earlier, we took an equally weighted mean of the models' quantile forecasts in the ensemble construction (see Section 2.2).If a model did not submit forecasts, the model's score was stored as 'NA'. When compiling the scores, the 'NA' values were processed in three ways: excluded from the analysis, substituted with the worst score for the combination of forecast date, location, and horizon, or substituted with the average score for the same combination. Here, we present the results obtained using the most conservative approach, in which 'NA' values were replaced with the worst scores. Results from other approaches show patterns similar to those observed below. The details are in the supplement (see Supplemental Section 4).

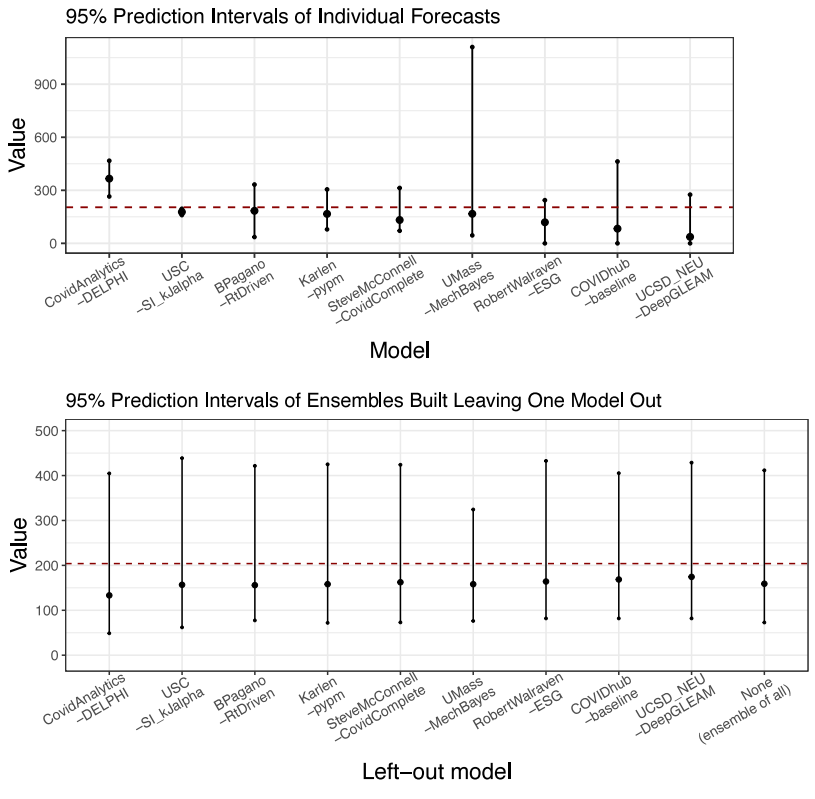
Overall, the importance metrics measured through the two computational algorithms were highly correlated in the positive direction with the negative WIS (Fig. 6). That is, on average, the more accurate a model was by −WIS, the more important a role it played in contributing to the accuracy of an ensemble.

In certain instances, the rankings of models by −WIS or by importance metrics differed. For example, the Karlen–pypm and BPagano–RtDriven models were the top two models by −WIS and by all importance metrics. Although BPagano-RtDriven showed higher accuracy by −WIS, Karlen-pypm showed greater importance on average, despite being substantially penalized for its missing values by assigning the worst score per the corresponding task for each metric, while BPagano-RtDriven was not penalized. This suggests that the Karlen-pypm model added more value than the BPagano-RtDriven model in its ability to contribute to ensemble predictions meaningfully. We also observe that USC-SI_kJalpha, which had a worse negative WIS, showed greater importance than MOBS-GLEAM_COVID (Table 1, Fig. 7), where the penalties incurred by both models were comparable. This implies that even models with low accuracy, as measured by −WIS, can provide a unique perspective that distinguishes them from other models as standalone predictive models and thereby further improve the average ensemble.

Factors not captured by −WIS but that influence the importance metric can be explained by model similarity. In Eq. (11), the importance metric is decomposed into individual forecast skills and the similarities of forecast errors from different component models for point forecasts.



(a)



(b)

Fig. 5. (5(a)) Model importance of each model versus negative WIS in Massachusetts on the target end date 2021-12-25. CovidAnalytics-DELPHI is the most important and also the least accurate by −WIS. (5(b)) Predictive medians and 95% Prediction intervals (PIs) of individual forecasts (top) and ensemble forecasts built leaving one model out (bottom) on target end date 2021-12-25. For example, the lines on the far left indicate PI for the CovidAnalytics-DELPHI model on the top panel and PI for the ensemble created without the CovidAnalytics-DELPHI model on the bottom panel. None(ensemble of all) represents an ensemble model built on all nine individual models. In each PI, the endpoints indicate 0.025 and 0.975 quantiles, and the mid-point represents the 0.5 quantile (predictive median). The horizontal dashed lines represent the eventual observation. The ensemble without CovidAnalytics-DELPHI is the only ensemble model with a point estimate below 150. The models on the *x*-axis are listed in order of model importance.

This concept can also be applied to probabilistic quantile-based forecasts, as demonstrated in Section 2.5.1. This similarity is understood in terms of how often the prediction errors fall "on the same side" of the observation and how much a particular model corrects errors from other models.

In general, the importance metrics for different computational algorithms (LASOMO/LOMO) are highly correlated with each other (Fig. 6). The relative ordering of models is not particularly sensitive to this choice. However, the importance metric calculated in LOMO, denoted

Table 1
Summary of negative WIS and importance metrics (Φ), sorted by −WIS. The number of predictions represents the total forecasts made by each model, with the percentage of the total number of predictions shown in parentheses, for the 50 US states across 1–4 week horizons from November 2020 to November 2022 (109 weeks). All scores were averaged across all forecast dates, locations, and horizons. In the importance metric notation (Φ), the superscript indicates the algorithm method; Φ^{lomo} represents the average importance metric based on leave one model out algorithm and Φ^{lasomo} represents the average importance metric based on leave all subsets of models out algorithm. The best value in each column is highlighted in bold.

Model	−WIS	Φ^{lasomo}	Φ^{lomo}	Number of predictions (%)
BPagano-RtDriven	−**40.2**	2.81	0.71	21 800 (100)
Karlen-pypm	−41.2	**3.11**	**0.92**	20 400 (93.6)
GT-DeepCOVID	−42.8	1.87	0.17	20 724 (95.1)
MOBS-GLEAM_COVID	−45.8	1.06	−0.21	20 596 (94.5)
CU-select	−47.3	1.64	0.24	21 000 (96.3)
RobertWalraven-ESG	−49.8	0.94	−0.09	19 992 (91.7)
USC-SI_kJalpha	−51.7	1.23	0.21	20 900 (95.9)
COVIDhub-baseline	−52.1	0.10	−0.62	21 800 (100)
UCSD_NEU-DeepGLEAM	−52.6	−0.13	−0.70	20 596 (94.5)
PSI-DRAFT	−71.7	−1.94	−1.00	19 988 (91.7)

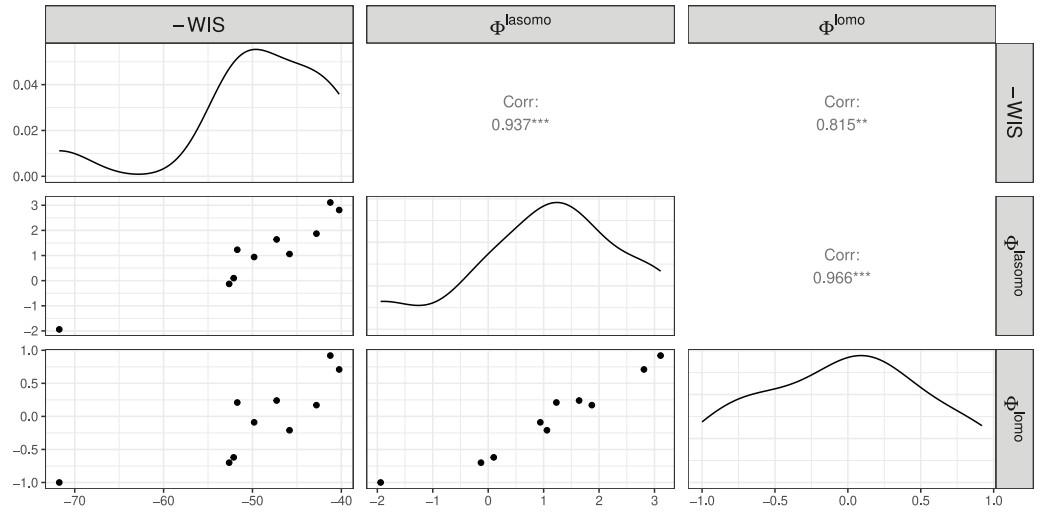


Fig. 6. Relationship between summary metrics computed across the entire evaluation period. In the importance metric notation (Φ), the superscript indicates the algorithm method; Φ^{lomo} represents the average importance metric based on leave one model out algorithm and Φ^{lasomo} represents the average importance metric based on leave all subsets of models out algorithm. One black dot corresponds to one model, with the position indicating the average scores across the entire evaluation period for the metrics in the row and column of the plot matrix.

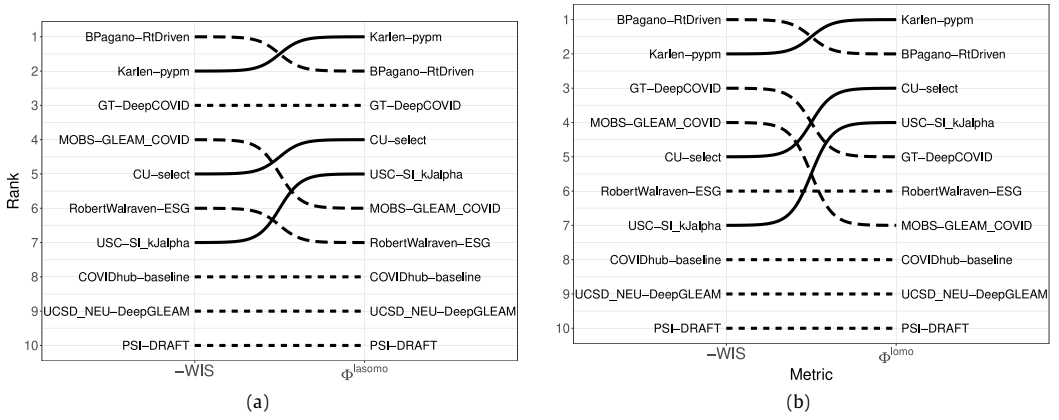


Fig. 7. Comparison of model ranks as measured by the negative WIS against different importance metrics: (a) −WIS vs. Φ^{lasomo} and (b) −WIS vs. Φ^{lomo}. Solid lines indicate cases where the importance metric rank is higher than the negative WIS rank, dashed lines indicate lower ranks, and dotted lines represent equal ranks.

by Φ^{lomo}, is consistently lower than the importance metric calculated in LASOMO, denoted by Φ^{lasomo}, for each model. Notably, many models that had positive scores in the LASOMO approach exhibit negative scores in the LOMO approach (Table 1). This can be interpreted as meaning that it is harder for a model to add value when all other models are already in the mix. It is because Φ^{lomo} represents the model's marginal contribution to the subset that includes all the other models, and it is considered only a part of Φ^{lasomo}.

As a complementary analysis, we also explored the variability of LOMO metrics across different subset sizes, which illustrates their influence on LASOMO metrics (see Supplement, Section 5). We found that subsets with a small number of models sometimes exhibit high variance in the LOMO metrics, potentially leading to instability in the LASOMO metrics. On the other hand, the LOMO model's importance scores for a given model are generally stable when one or two other models are removed from the pool.

4. Discussion

We have developed an importance metric based on the Shapley value concept. While earlier studies applied this concept to measure the contribution of individual predictors of a predictive model (Borup et al., 2024; Giudici & Raffinetti, 2021) or ensemble component models (Adiga et al., 2023; Pompigna & Rupi, 2018) in terms of point prediction accuracy, we have explored the concept in the context of probabilistic ensemble forecasts of epidemics. We also provided a detailed understanding of the accuracy and similarity decomposition in the model importance metric, revealing the conditions under which a component model is rewarded in the presence of other models. This highlights a key distinction from the ambiguity decomposition proposed by Brown et al. (2005).

The Weighted Interval Score (WIS), a proper scoring rule designed for quantile forecasts, was utilized by the US COVID-19 Forecast Hub and several other collaborative forecasting challenges (Howerton et al., 2023; Mathis, et al., 2024; Sherratt, et al., 2023). While WIS is effective in measuring each model's performance independently, it does not offer a complete picture of a model's contribution in an ensemble setting. The importance metric approaches we introduce in this work provide insights that WIS cannot capture, as it relies on predictions from other models. No matter how accurate a prediction is, if its prediction errors are highly similar to those of other models, its impact on the ensemble may not be as great as that of a model that has lower accuracy but offsets the errors of other models. This aspect of importance metrics is especially relevant for hub organizations like the US CDC, which collect forecasts from a variety of models and combine them to generate ensemble forecasts for communication with the public and decision-makers (Fox et al., 2024).

We proposed two algorithms for assessing model importance: leave-one-model-out (LOMO) and leave-all-subsets-of-models-out (LASOMO). In the LASOMO algorithm, we used permutation-based weights to account for how a model's contribution can vary with ensemble size, thereby distinguishing our work from that of Adiga et al. (2023), who use an equal-weighting scheme. Notably, LOMO is a special case of LASOMO, in which only a single subset is considered, namely, all models except the target component model. This makes LOMO simple and easy to implement and significantly more efficient than LASOMO, especially when dealing with many component models. However, when the number of component models is relatively small (e.g., fewer than 10), LASOMO becomes computationally feasible and may be preferred, as it provides a more comprehensive evaluation by considering all possible combinations of component models.

This study has several limitations. While we used the widely adopted mean ensemble method in the application, it is often vulnerable to individual forecasters with outlying or poorly calibrated predictions, which can increase forecast uncertainty or decrease overall ensemble reliability. Additionally, our use of Shapley values, while providing insights, was constrained by underlying assumptions, such as assigning the empty set a zero value in the characteristic function, that were not fully met in our setting. Thus, the obtained values may not precisely represent Shapley values but rather provide an informal approximation. The computational cost of implementing the LASOMO algorithm is also a challenge. As the number of models increases, computational time grows exponentially because 2^n subsets must be considered for n models in the Shapley value calculation. Furthermore, it is almost impossible to have all models consistently submit their predictions over a given period when there are many participating models, so it is inevitable to have missing data for unsubmitted predictions. Consequently, the Shapley value can be unstable or misleading, as it is highly sensitive to such missingness. Because the choice of how to handle these missing values during scoring can lead to variations in the resulting importance metrics and rankings of the component models, careful consideration is required when selecting a handling method. Further exploration of this issue is needed for comprehensive guidelines.

We also envision several directions for future research. A naïve forecast could serve as a baseline model, replacing the forecast associated with the empty set of component models, which we excluded in this study. Another potential direction is to explore the application of model importance measures in the context of ensemble forecasts that assign weights to individual component models. In this case, more deliberate strategies should be explored to account for the different levels of consistency and reliability across models in the weighting scheme (Ray et al., 2023). Moreover, the components of the LASOMO metric computed using subsets of a few models sometimes exhibit high variance. While this is less of an issue for LOMO metrics when a reasonable number of models are available, several approaches could be explored to reduce the impact of high-variance metrics and outliers. We could refine the LASOMO framework to develop variance-adjusted versions of the metric. It could also be valuable

to explore ways to extend the Rank Graduation Accuracy (RGA) metrics used in the point predictions (Giudici & Raffinetti, 2025) in the context of a probabilistic forecasting setting. Such a probabilistic implementation of RGA could improve robustness, as it would be less affected by outliers.

5. Conclusions

Despite the rising popularity of ensemble models, there is currently a lack of comprehensive evaluation metrics to assess the individual contributions of each model. Traditional practice involves setting up a leaderboard to independently evaluate the accuracy and effectiveness of individual prediction models using appropriate scoring rules. Our proposed importance metric addresses this gap by providing a novel, distinctive metric for assessing the role of each model within the ensemble, adding a unique dimension to the assessment of forecasting models.

This paper presents a decomposition of the model importance metric, which mathematically demonstrates how an individual model's accuracy and its interactions with other component models influence the measure. Simulation studies support this theoretical framework. In a case study, its application is illustrated in a real-world setting. These analyses provide both formal and intuitive explanations of the realized values of the model importance metrics and highlight how the model importance metrics can be used to understand how individual models have improved or degraded predictive accuracy. An extensive application further highlights the relationship between a widely used accuracy metric and our model importance metric.

The implication of this work is that our proposed importance metric provides novel insights, offering new information beyond traditional accuracy metrics. Our method provides a solid theoretical basis and clear criteria for quantifying a component model's contribution to ensemble performance. Moreover, leveraging the importance metric can incentivize original modeling approaches, thereby fostering a diverse landscape of perspectives among modelers and ultimately enriching the forecasting ecosystem.

CRediT authorship contribution statement

Minsu Kim: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis. **Evan L. Ray:** Writing – review & editing, Supervision, Resources, Methodology, Conceptualization. **Nicholas G. Reich:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Funding

This work has been supported by the National Institutes of General Medical Sciences (R35GM119582) and the US CDC (1U01IP001122). The content is solely the responsibility of the authors and does not necessarily represent the official views of NIGMS, the National Institutes of Health, or CDC.

Acknowledgment

We acknowledge Daniel Sheldon for helping seed the original idea for this work. This study started by taking the kernel of his ideas about having models ranked based on their unique contribution to making the ensemble better. Conversations with Mark Wilson were also helpful in the nascent stages of thinking about and understanding Shapley values.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2025.12.006.

Data and code availability

The code and data used to reproduce the results in this study are archived on Zenodo (https://doi.org/10.5281/zenodo.17954018) and are also available on GitHub at https://github.com/mkim425/replication_model-importance.

References

Adiga, A., Kaur, G., Wang, L., Hurt, B., Porebski, P., Venkatramanan, S., Lewis, B., & Marathe, M. V. (2023). Phase-informed Bayesian ensemble models improve performance of COVID-19 forecasts. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(13), 15647–15653.

Batchelor, R., & Dua, P. (1995). Forecaster diversity and the benefits of combining forecasts. *Management Science*, *41*(1), 68–75.

Borup, D., Goulet Coulombe, P., Rapach, D., Schütte, E. C. M., & Schwenk-Nebbe, S. (2024). The anatomy of out-of-sample forecasting accuracy. FRB Atlanta Working Paper . No. 2022–16B.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: a survey and categorisation. *Information Fusion*, *6*, 5–20.

Centers for Disease Control and Prevention (2023). COVID-19 forecasting and mathematical modeling. (Accessed 25 October 2024).

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

Cramer, E. Y., Huang, Y., Wang, Y., Ray, E. L., Cornell, M., Bracher, J., Brennen, A., Rivadeneira, A. J. C., Gerding, A., House, K., Jayawardena, D., Kanji, A. H., Khandelwal, A., Le, K., Mody, V., Mody, V., Niemi, J., Stark, A., Shah, A., ... U. S. COVID-19 Forecast Hub Consortium (2022). The United States COVID-19 forecast hub dataset. *Scientific Data*, *9*(1), 462.

Cramer, E. Y., Ray, E. L., Lopez, V. K., Bracher, J., Brennen, A., Rivadeneira, A. J. C., Gerding, A., Gneiting, T., House, K. H., Huang, Y., Jayawardena, D., Kanji, A. H., Khandelwal, A., Le, K., Mühlemann, A., Niemi, J., Shah, A., Stark, A., Wang, Y., ... Reich, N. G. (2022). Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences*, *119*(15), Article e2113561119.

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, *20*(5), 533–534.

Fox, S. J., Kim, M., Meyers, L. A., Reich, N. G., & Ray, E. L. (2024). Optimizing disease outbreak forecast ensembles. *Emerging Infectious Diseases*, *30*, 1967–1969.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.
Giudici, P., & Raffinetti, E. (2021). Shapley-lorenz explainable artificial intelligence. *Expert Systems with Applications*, *167*, Article 114104.
Giudici, P., & Raffinetti, E. (2025). RGA: a unified measure of predictive accuracy. *Advances in Data Analysis and Classification*, *19*, 67–93.
Gneiting, T., & Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, *310*(5746), 248–249.
Howerton, E., Contamin, L., Mullany, L. C., Qin, M., Reich, N. G., Bents, S., Borchering, R. K., et al. (2023). Evaluation of the US COVID-19 scenario modeling hub for informing pandemic response under uncertainty. *Nature Communications*, *14*, 7260.
Kang, Y., Cao, W., Petropoulos, F., & Li, F. (2022). Forecast with forecasts: Diversity matters. *European Journal of Operational Research*, *301*(1), 180–190.
Lamberson, P. J., & Page, S. E. (2012). Optimal forecasting groups. *Management Science*, *58*(4), 805–810.
Lee, J.-Y., Marotzke, J., Bala, G., Cao, L., Corti, S., Dunne, J., Engelbrecht, F., Fischer, E., Fyfe, J., Jones, C., Maycock, A., Mutemi, J., Ndiaye, O., Panickal, S., & Zhou, T. (2021). Future global climate: Scenario-based projections and near-term information. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu, & B. Zhou (Eds.), *Climate change 2021: the physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change* (pp. 553–672). Cambridge University Press.
Lichtendahl, K. C., & Winkler, R. L. (2020). Why do some combinations perform better than others? *International Journal of Forecasting*, *36*(1), 142–149.
Lutz, C. S., Huynh, M. P., Schroeder, M., Anyatonwu, S., Dahlgren, F. S., Danyluk, G., Fernandez, D., Greene, S. K., Kipshidze, N., Liu, L., Mgbere, O., McHugh, L. A., Myers, J. F., Siniscalchi, A., Sullivan, A. D., West, N., Johansson, M. A., & Biggerstaff, M. (2019). Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. *BMC Public Health*, *19*(1), 1659.
Mathis, S. M., Webber, A. E., León, T. M., Murray, E. L., Sun, M., White, L. A., Brooks, L. C., Green, A., Hu, A. J., Rosenfeld, R., Shemetov, D., Tibshirani, R. J., McDonald, D. J., Kandula, S., Pei, S., Yaari, R., Yamana, T. K., Shaman, J., Agarwal, P., ... Borchering, R. K. (2024). Title evaluation of FluSight influenza forecasting in the 2021–22 and 2022–23 seasons with a new target laboratory-confirmed influenza hospitalizations. *Nature Communications*, *15*(1), 6289.
Pompigna, A., & Rupi, F. (2018). Comparing practice-ready forecast models for weekly and monthly fluctuations of average daily traffic and enhancing accuracy by weighting methods. *Journal of Traffic and Transportation Engineering (English Edition)*, *5*, 239–253.
Ray, E. L., Brooks, L. C., Bien, J., Biggerstaff, M., Bosse, N. I., Bracher, J., Cramer, E. Y., Funk, S., Gerding, A., Johansson, M. A., Rumack, A., Wang, Y., Zorn, M., Tibshirani, R. J., & Reich, N. G. (2023). Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States. *International Journal of Forecasting*, *39*(3), 1366–1383.
Shapley, L. S. (1953). A value for n-person games. *Contribution To the Theory of Games*, *2*.
Sherratt, K., Gruson, H., Grah, R., Johnson, H., Niehus, R., Prasse, B., Sandmann, F., Deuschel, J., Wolffram, D., Abbott, S., Ullrich, A., Gibson, G., Ray, E. L., Reich, N. G., Sheldon, D., Wang, Y., Wattanachit, N., Wang, L., Trnka, J., ... Funk, S. (2023). Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations. *ELife*, *12*, Article e81916.
Thomson, M. E., Pollock, A. C., Önkal, D., & Gönül, M. S. (2019). Combining forecasts: Performance and coherence. *International Journal of Forecasting*, *35*(2), 474–484.
Timmermann, A. (2006). Forecast combinations. In *Handbook of Economic Forecasting*: *Vol. 1*, (pp. 135–196).
Viboud, C., Sun, K., Gaffey, R., Ajelli, M., Fumanelli, L., Merler, S., Zhang, Q., Chowell, G., Simonsen, L., & Vespignani, A. (2018). The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics*, *22*, 13–21.