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

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

Abstract

Ensemble forecasting is generally recognized for its ability to outperform individual standalone models in infectious disease forecasting and is therefore considered a robust option for public health decision making and policy planning. The US COVID-19 Forecast Hub has produced a probabilistic ensemble forecast model of COVID-19 cases, hospitalizations, and deaths in the United States based on forecasts from individual models developed by more than 90 different research groups. As the Forecast Hub served as the official short-term forecast of the US Centers for Disease Control and Prevention, it is important to understand the relative importance and contributions of individual models to creating a highly accurate forecast combination. In this work, we propose two practical methods for evaluating the contribution of individual component models. One method uses a leave-one-model-out algorithm when building an ensemble and the other, based on the Shapley value in game theory, considers ensemble models constructed from all possible subsets of individual models. We aim to identify and evaluate methods for measuring the contributions of individual component models to ensemble accuracy. We explore how these metrics are related to the weighted interval score (WIS), a commonly used proper scoring rule for quantile forecasts, and illustrate how these methods provide distinct perspectives when evaluating how much value a component model adds to a probabilistic ensemble model in the presence of other models. Our results show that the most accurate model according to WIS does not always add the most value to the ensemble. This indicates that our proposed methods can be used to capture the contribution of individual models to a more accurate ensemble model, which is difficult to ascertain from standard accuracy metrics alone. This study offers valuable insights into understanding individual forecasting models' unique features and their roles in contributing to an ensemble model for a specific prediction task.

1 Introduction

2 Methods

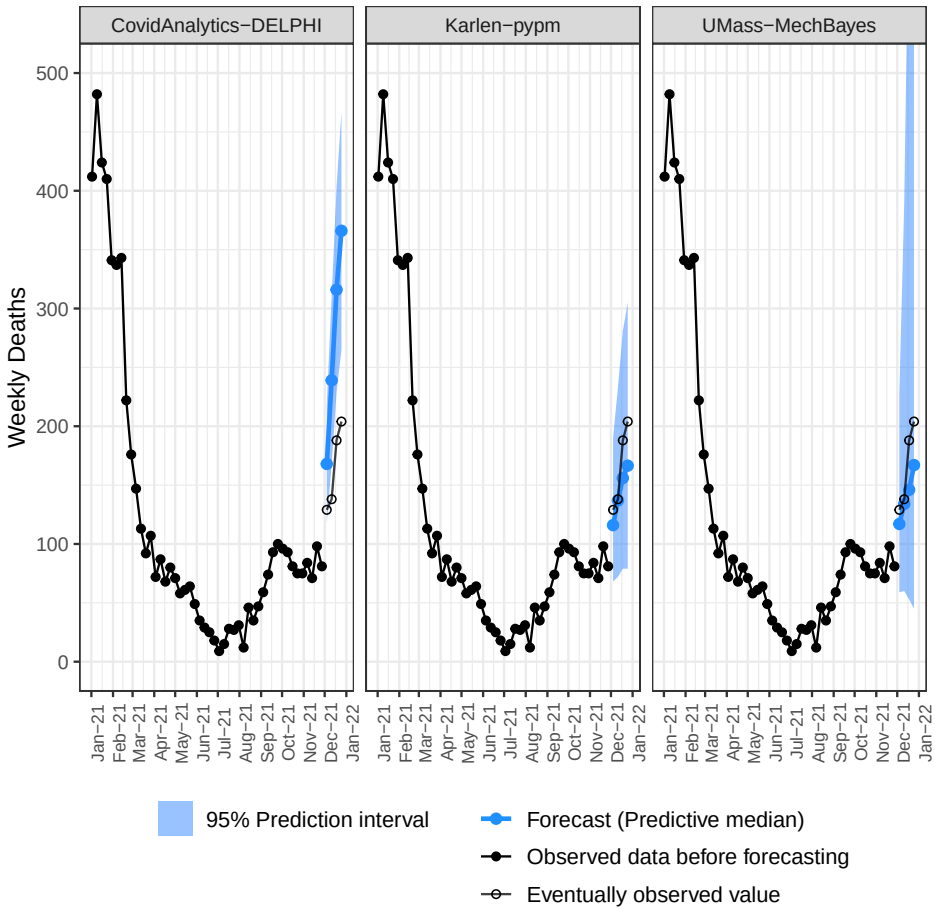

Figure 1: Forecasts of COVID-19 incident deaths at 1- through 4-week horizons in Massachusetts made on November 27, 2021 by three models. Black dots show historical data available as of November 28. Blue dots indicate predictive medians and the shaded bands represent 95% prediction intervals. The open black circles are observations 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.

3 Results

3.1 Application of the model importance metrics to forecast data from the US COVID-19 Forecast Hub

3.1.1 Case study: Relationship between importance score and WIS with data for deaths in Massachusetts in 2021

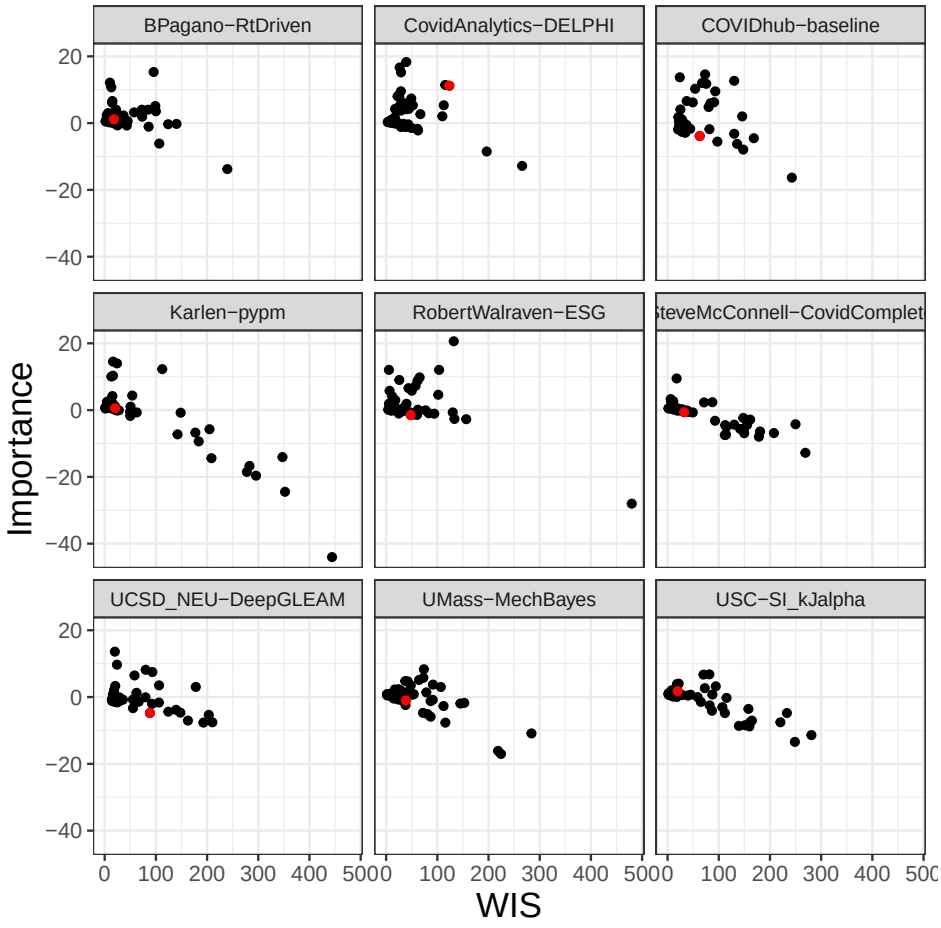

Figure 2: Model importance versus WIS by model for all weeks in 2021. Each point represents a pair of WIS and importance score evaluated at a certain week, and each panel contains 52 points accordingly. Red dots represent WIS and importance score pairs evaluated on December 25, 2021. The importance of an individual model as an ensemble member tends to be inversely correlated with that model's overall prediction accuracy.

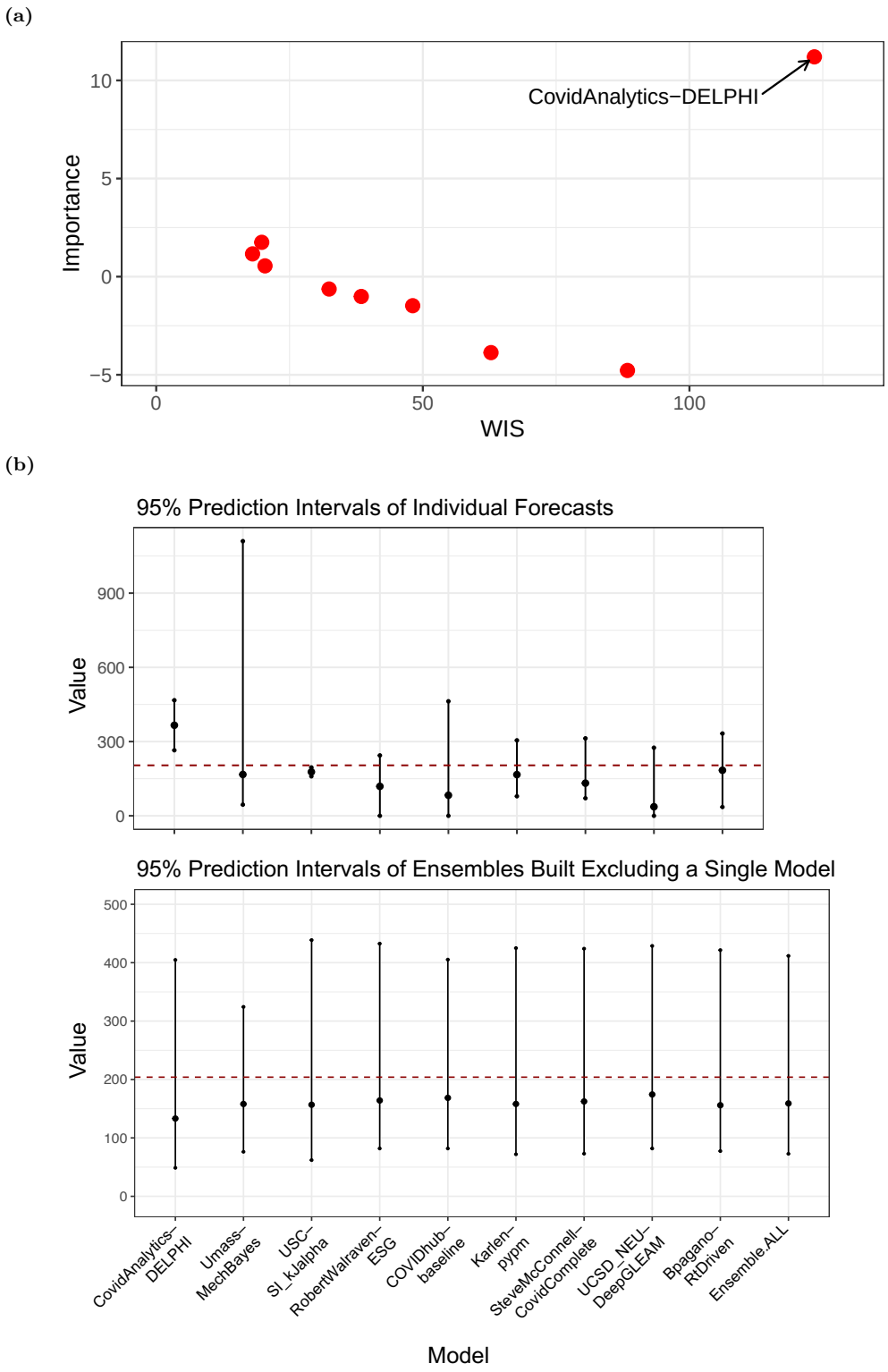

Figure 3: (b) 95% Prediction intervals (PIs) of individual forecasts (top) and ensemble forecasts built excluding a single model (bottom) on target end date 2021-12-25. For example, the lines on the far left indicate PI for the CovidAnalytics-DELPHI model on the top panel and PI for the ensemble created without the CovidAnalytics-DELPHI model on the bottom panel. Ensemble.ALL represents an ensemble model built on all nine individual models. In each PI, the end points 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 ensemble without UMass-MechBayes has the lowest dispersion among the ensemble models. (a) Model importance of each model versus WIS on target end date 2021-12-25. CovidAnalytics-DELPHI is the most important and also the least accurate by WIS.

3.1.2 Differences in importance scores measured by different algorithms and ensemble methods

Model	Submission rate (%)	Number of predictions (total:21800)
BPagano-RtDriven	100.0	21800
COVIDhub-baseline	100.0	21800
CU-select	96.3	21000
USC-SI_kJalpha	95.9	20900
GT-DeepCOVID	95.1	20724
MOBS-GLEAM_COVID	94.5	20596
UCSD_NEU-DeepGLEAM	94.5	20596
Karlen-pypm	93.6	20400
PSI-DRAFT	91.7	19988
RobertWalraven-ESG	91.7	19992

Table 1: Submission rates (rounded to one decimal place) and the number of individual forecasts made by 10 models for the US 50 states and 1-4 horizons from November 2020 to November 2022 (109 weeks).

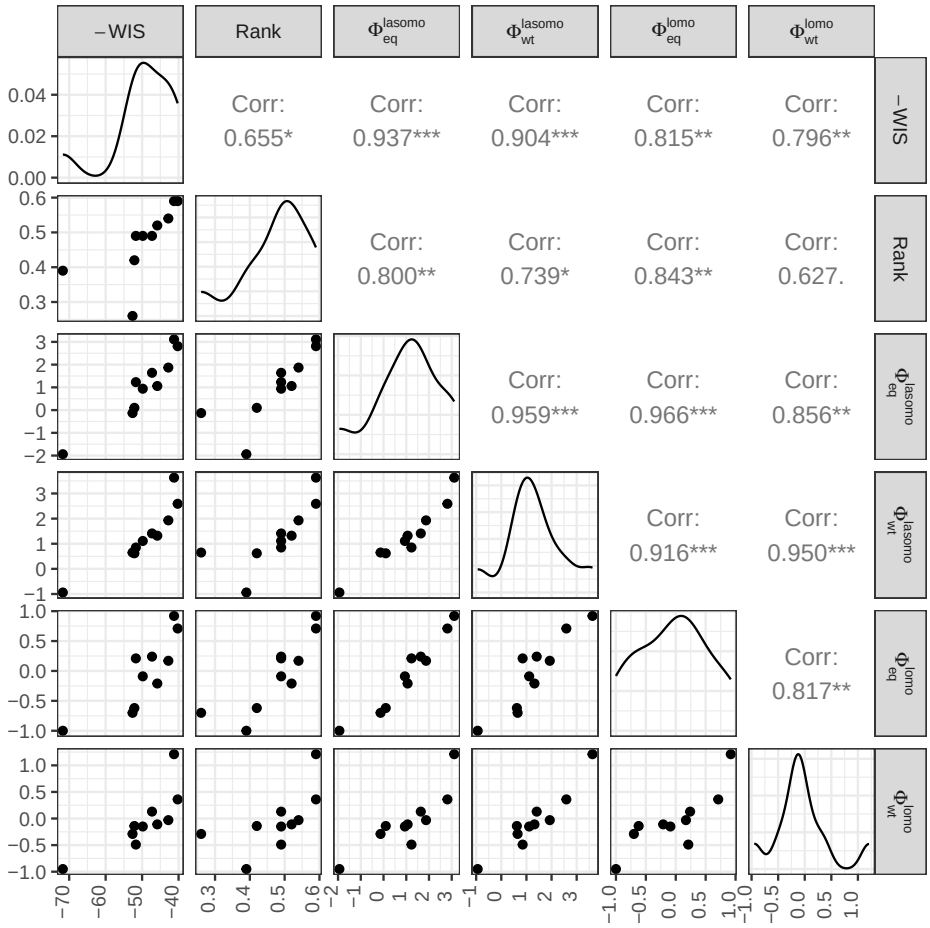

Figure 4: Relationship between summary metrics computed across the entire evaluation period. One black dot corresponds to one model, with the position representing the average scores across the entire evaluation period for the metrics corresponding to the row and column of the plot matrix.

4 Simulation Studies

4.1 Relationship between a component forecaster's bias and importance

Model	WIS	relWIS	Rank	$\Phi_{\text{eq}}^{\text{lasomo}}$	$\Phi_{\text{wt}}^{\text{lasomo}}$	$\Phi_{\text{eq}}^{\text{lomo}}$	$\Phi_{\text{wt}}^{\text{lomo}}$
BPagano-RtDriven	40.2	0.77	0.59	2.81	2.59	0.71	0.36
Karlen-pypm	41.2	0.79	0.59	3.11	3.63	0.92	1.21
GT-DeepCOVID	42.8	0.82	0.54	1.87	1.93	0.17	-0.03
MOBS-GLEAM_COVID	45.8	0.88	0.52	1.06	1.32	-0.21	-0.11
CU-select	47.3	0.91	0.49	1.64	1.41	0.24	0.13
RobertWalraven-ESG	49.8	0.96	0.49	0.94	1.11	-0.09	-0.15
USC-SI_kJalpha	51.7	0.99	0.49	1.23	0.85	0.21	-0.49
COVIDhub-baseline	52.1	1.00	0.42	0.10	0.62	-0.62	-0.14
UCSD_NEU-DeepGLEAM	52.6	1.01	0.26	-0.13	0.65	-0.70	-0.29
PSI-DRAFT	71.7	1.38	0.39	-1.94	-0.94	-1.00	-0.95

Table 2: Summary of WIS, relative WIS compared to the COVIDhub-baseline, standardized rank score, importance scores (Φ), sorted by mean WIS. All scores were averaged across all forecast dates, locations, and horizons. In the importance score notation (Φ), the superscript indicates the algorithm method, and the subscript indicates the ensemble method used. For example, $\Phi_{\text{wt}}^{\text{lomo}}$ represents the mean important score based on leave one model out algorithm with weighted (trained) ensemble and $\Phi_{\text{eq}}^{\text{lasomo}}$ represents the mean important score based on leave all subsets of models out algorithm with equally weighted (untrained) ensemble. The best value in each column is highlighted in bold.

4.2 Relationship between component forecaster dispersion and importance

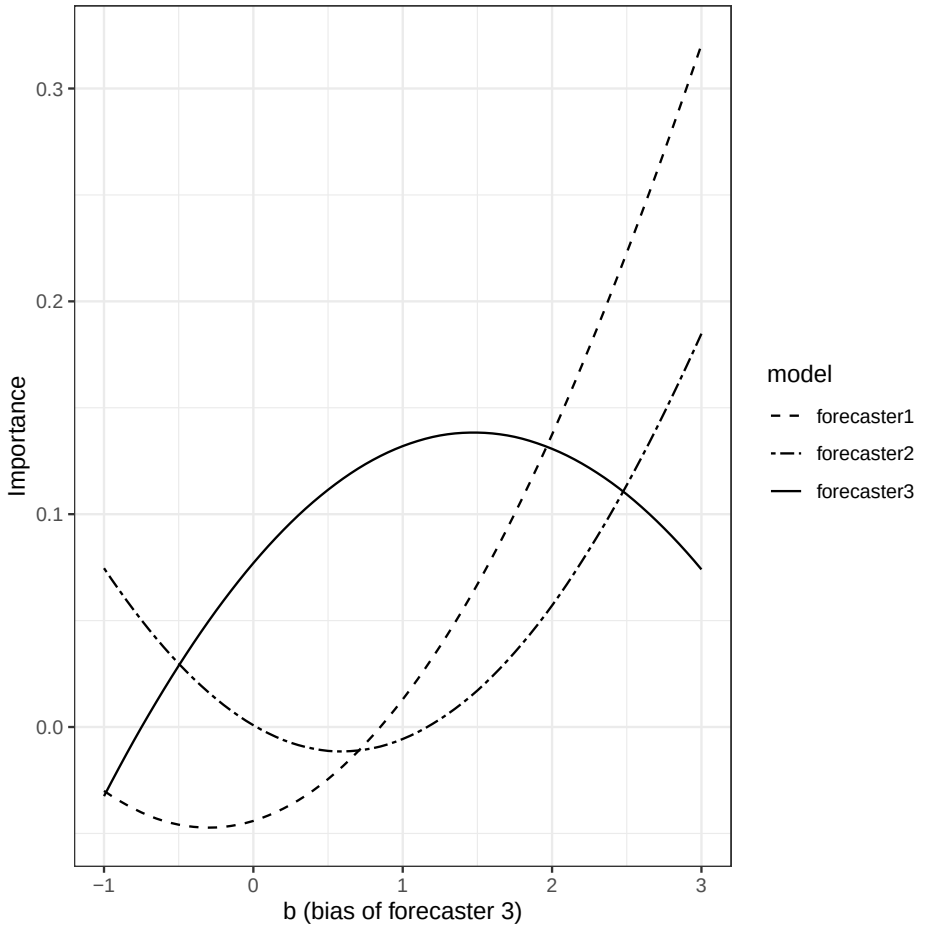

Figure 5: Importance of three forecasters by various biases of forecaster 3. $F_{1,t} = \mathrm{N}(-1, 1)$, $F_{2,t} = \mathrm{N}(-0.5, 1)$, and $F_{3,t} = \mathrm{N}(b, 1)$. Importance scores were calculated and averaged over 1000 replicates of the forecasting experiments conducted at each value of b, incremented by 0.05 from -1 to 3. The results are displayed in different line patterns by model (dashed line for forecaster 1, dash-dotted line for forecaster 2, and solid line for forecaster 3).

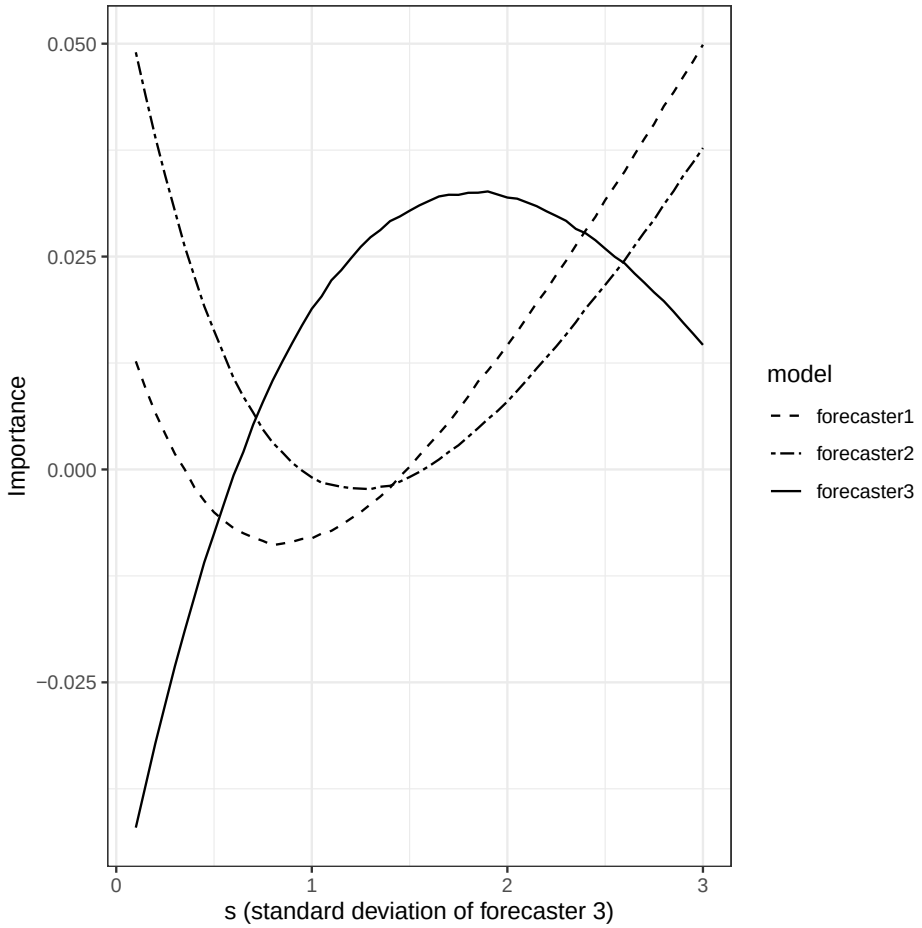

Figure 6: Importance of three forecasters as a function of dispersion of forecaster 3. $F_{1,t} = \mathrm{N}(0, 0.5^2)$, $F_{2,t} = \mathrm{N}(0, 0.7^2)$, and $F_{3,t} = \mathrm{N}(0, s^2)$. Importance scores were calculated and averaged over 1000 replicates of the forecasting experiments conducted at each value of s, incremented by 0.05 from 0.1 to 3. The results are displayed in different line patterns by model (dashed line for forecaster 1, dash-dotted line for forecaster 2, and solid line for forecaster 3).