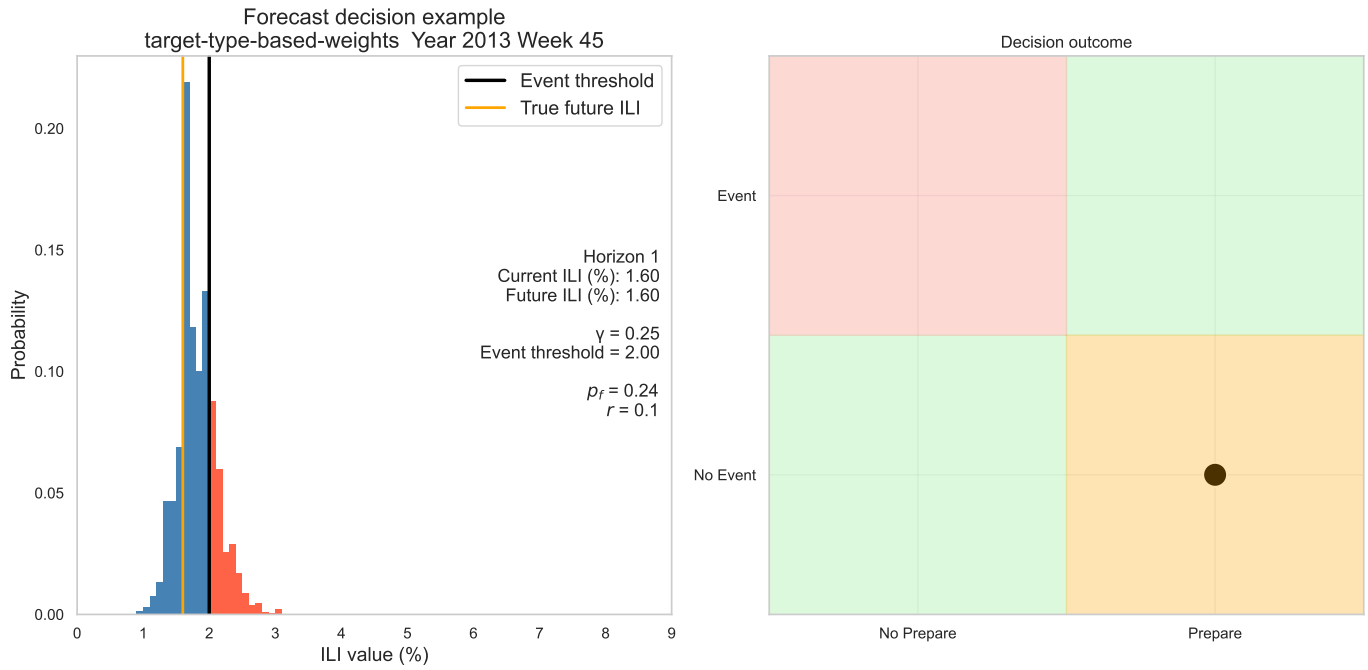

Evaluating the Relationship Between Forecast Accuracy and Economic Value of Forecast

A Cost–Loss Based Evaluation of Influenza Forecasts

Master's thesis in Complex Adaptive Systems

ALBIN BLOMSTER

DEPARTMENT OF MATHEMATICAL SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's thesis 2026

Evaluating the Relationship Between Forecast Accuracy and Economic Value of Forecast

A Cost–Loss Based Evaluation of Influenza Forecasts

ALBIN BLOMSTER

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
Chalmers University of Technology
Gothenburg, Sweden 2026

Evaluating the Relationship Between Forecast Accuracy and Economic Value of Forecast
A Cost–Loss Based Evaluation of Influenza Forecasts
ALBIN BLOMSTER

Supervisor & Examiner: Phillip Gerlee, Department of Mathematical Sciences

Master's Thesis 2026
Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Visualization of the cost–loss framework used to evaluate the practical value of probabilistic influenza forecasts. The figure illustrates how forecast probabilities are converted into preparedness decisions under uncertainty.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

Evaluating the Relationship Between Forecast Accuracy and Value of Forecast
A Cost–Loss Based Evaluation Using Influenza Forecasts
Albin Blomster
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

Epidemiological forecast evaluation commonly focuses on statistical accuracy measures such as Mean Absolute Error (MAE), Weighted Interval Score (WIS), and Log Score (LS). However, accurate forecasts do not necessarily translate into improved decision-making performance. Therefore, this thesis investigates the relationship between forecast accuracy and the value of a forecast within probabilistic influenza forecasting.

The study uses data from the CDC FluSight forecasting challenge, where forecasting models provide probabilistic predictions of future Influenza-like Illness (ILI) prevalence. Event-based forecasting scenarios are constructed through threshold definitions based on relative growth in ILI prevalence. Forecast probabilities are obtained from predictive distributions and used within a cost–loss decision framework to determine preventive actions.

Ten forecasting models are evaluated across multiple event definitions, cost–loss ratios, and forecast horizons. Decision performance is quantified using value scores and integrated value measures. Furthermore, Spearman rank correlation analysis is used to investigate the relationship between forecast accuracy rankings and decision-based rankings.

The results indicate that forecast accuracy and value score rankings are generally positively correlated across event definitions, cost–loss ratios, and forecast horizons, although the relationship is not perfectly aligned. While the ensemble model "target-type-based-weights" and the model "Reichlab-KCDE" perform strongly under both evaluation frameworks, differences between statistical and decision-based rankings are still observed. These findings suggest that traditional forecast accuracy measures and decision-based approaches provide complementary perspectives for evaluating the practical usefulness of probabilistic forecasts.

Keywords: Influenza Forecasting, Probabilistic Forecasting, Value of Forecast, Cost—Loss Analysis, Decision Theory, FluSight, Forecast Evaluation, Rank Correlation.

Acknowledgements

I would like to express my sincere gratitude to my supervisor and examiner, Philip Gerlee, for valuable guidance, insightful feedback, and continuous support throughout this thesis project. Your ability to explain concepts clearly and provide direction during challenging parts of the process has been greatly appreciated. I am also especially grateful for your understanding, encouragement, and patience throughout the project. Our discussions and your advice have played an important role in shaping both this work and the overall experience of writing the thesis.

I would also like to thank my friends, who throughout the thesis period made sure that lunch at Kårrestaurangen remained a recurring and highly appreciated break from writing. Especially since the thesis was written during the spring in Sweden, taking a step away from the thesis to sit outside in the sun, enjoy good conversations, and exchange some banter became a valuable part of the routine. Besides providing a welcome break, these moments occasionally also resulted in useful discussions and inspiration for the thesis.

A special thanks also goes to my family for their continuous encouragement and support throughout my studies. Even if the details of probabilistic influenza forecasting and value of forecast may not always have been entirely clear, you always made sure I stayed on track and kept moving toward the finish line.

Finally, I would like to thank the silent study room next to Laplace for faithfully serving as my unofficial office throughout this process. Despite technically being available to everyone, its consistency, availability, and quiet company did not go unnoticed.

Albin Blomster, Gothenburg, May 20 2026

List of Acronyms

Below is the list of acronyms used throughout this thesis.

ILI	Influenza-like Illness
CDC	Centers for Disease Control and Prevention
WIS	Weighted Interval Score
MAE	Mean Absolute Error
LS	Log Score
VS	Value Score
IV	Integrated Value

Nomenclature

Below is the nomenclature of indices, parameters, and variables used throughout this thesis.

Indices

t	Current week index
h	Forecast horizon index
i	Forecast instance index
m	Forecast model index

Parameters

γ	Event definition parameter (growth threshold)
r	Cost–loss ratio C/L
C	Preventive action cost
L	Loss incurred if no action is taken and the event occurs
N	Total number of forecast instances

Variables

Y_t	Current observed ILI value
Y_{t+h}	Future observed ILI value at horizon h
p_f	Forecast event probability
p_b	Historical baseline probability
$E^{(i)}$	Binary event indicator
E_f	Forecast-based loss
E_b	Baseline loss

E_p	Perfect foresight loss
VS	Value score
IV	Integrated value score

Contents

List of Acronyms **ix**

Nomenclature **x**

List of Figures **xv**

List of Tables **xix**

1 Introduction **1**
1.1 Aim and Research Questions 3

2 Theory **5**
2.1 Epidemiological Forecasting 5
2.2 Probabilistic forecasts and the FluSight challenge 6
2.3 Forecast Accuracy Evaluation 7
2.4 Event scenarios 9
2.5 Cost–loss scenario 9
2.6 Value of Forecast 10
2.7 Rank correlation 11

3 Methods **13**
3.1 FluSight Data Repository 13
3.2 Model Selection 15
3.3 Data Preparation 16
3.4 Forecast Accuracy Evaluation 16
3.5 Event Construction 17
3.6 Cost–loss implementation 18
3.7 Value of Forecast Computation 19
3.8 Integrated Value Measures 21
3.9 Rank Correlation Analysis 22

4 Results **23**
4.1 Event and Data Characteristics 23
4.1.1 Seasonal ILI trajectories 23
4.1.2 Growth rate distribution 24
4.1.3 Event frequency by horizon 25
4.1.4 Event probability by season 26

4.2 Baseline Behavior 26
4.2.1 Historical belief 27
4.3 Forecast Accuracy 27
4.3.1 Accuracy performance by horizon 27
4.4 Value of Forecast 29
4.4.1 Value score curve 29
4.4.2 Integrated decision value 30
4.4.3 Decision value surface 31
4.5 Rank Correlation 33

5 Discussion **35**
5.1 Forecast Accuracy vs Decision Value 35
5.2 Effect of Event Severity and Horizon 36
5.3 Interpretation of the Cost–Loss Framework 37
5.4 Baseline Construction and Historical Information 38
5.5 Limitations 38
5.6 Future Research 39

6 Conclusion **41**

Bibliography **43**

A Appendix 1 **I**

B Appendix 2 **XXI**

List of Figures

2.1 Illustrative example of probabilistic ILI forecasting. Historical observations are available up to the forecast origin, after which predictive distributions are generated 1 and 4 weeks ahead. The realized ILI values are observed only after the target weeks occur and can then be compared with the forecast distributions. 7

3.1 Example of the cost–loss decision framework for a single forecast instance with event definition $\gamma = 0.25$, forecast horizon $h = 1$, and cost–loss ratio $r = 0.10$. The left panel shows the probabilistic forecast together with the event threshold used to compute the forecast probability p_f, while the right panel displays the corresponding decision outcome in the cost–loss matrix. 18

4.1 Observed ILI prevalence across influenza seasons as a function of *season time*. Each line corresponds to one influenza season. 23
4.2 Empirical distribution of the growth rate $\frac{Y_{t+h}}{Y_t}$ across forecast horizons $h = 1, 2, 3, 4$. 24
4.3 Event frequency as a function of the event definition parameter γ across forecast horizons $h = 1, 2, 3, 4$. 25
4.4 Empirical event frequencies across influenza seasons as a function of the event definition parameter γ for different forecast horizons. 26
4.5 Comparison between the true event frequency, the historical baseline probability p_b, the static baseline probability, and the decision threshold $r = 0.19$ for $\gamma = 0.1$ and forecast horizon $h = 1$. 27
4.6 Forecast accuracy metrics MAE, LS, and WIS across forecasting models and forecast horizons $h = 1, 2, 3, 4$. 28
4.7 Value score as a function of the cost–loss ratio r for the ensemble model "target-type-based-weights" with $\gamma = 0.1$ and forecast horizon $h = 4$. The dashed line corresponds to the break-even value $VS = 0$. . 29
4.8 Integrated value scores across event definitions γ for all forecasting models and forecast horizons $h = 1, 2, 3, 4$. The dashed line corresponds to the break-even value $IV = 0$. 30
4.9 Value score surfaces for the ensemble model "target-type-based-weights" across forecast horizons $h = 1, 2, 3, 4$. The heatmaps display the value score as a function of the cost–loss ratio r and the event definition parameter γ. 31

4.10 Value score surfaces for the "CUBMA" model across forecast horizons $h = 1, 2, 3, 4$. The heatmaps display the value score as a function of the cost–loss ratio r and the event definition parameter γ. 32
4.11 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.20$ across forecast horizons and cost–loss ratios. 33
4.12 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.45$ across forecast horizons and cost–loss ratios. 34

A.1 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.00$. II
A.2 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.03$. III
A.3 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.05$. IV
A.4 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.08$. V
A.5 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.10$. VI
A.6 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.12$. VII
A.7 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.15$. VIII
A.8 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.18$. IX
A.9 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.23$. X
A.10 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.25$. XI
A.11 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.28$. XII
A.12 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.30$. XIII
A.13 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.33$. XIV
A.14 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.35$. XV
A.15 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.38$. XVI
A.16 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.40$. XVII
A.17 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.43$. XVIII
A.18 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.48$. XIX

A.19 Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.50$. XX

B.1 Example of the cost–loss decision framework where preventive action is taken, and the event occurs. The left panel shows the probabilistic forecast together with the event threshold used to compute the forecast probability p_f, while the right panel displays the corresponding decision outcome in the cost–loss matrix. XXI
B.2 Example of the cost–loss decision framework where no preventive action is taken, and the event occurs. XXII
B.3 Example of the cost–loss decision framework where no preventive action is taken, and the event does not occur. XXIII

List of Tables

2.1 Cost–loss matrix for a binary decision problem. C denotes the cost of preventive action and L the loss incurred if no action is taken and the event occurs. 10

3.1 Forecasting models included in the analysis. 15

1

Introduction

Epidemiological forecasting has emerged as a response to the need for informed decision-making in public health under uncertainty. During infectious disease outbreaks, decision-makers such as public-health agencies and hospital leadership must take time-sensitive decisions regarding, for example, healthcare capacity, staffing, and preparedness measures. These decision-makers often possess detailed contextual knowledge. For example, hospital administrators typically have a strong understanding of operational constraints and consequences but may lack the expertise required to formulate and interpret epidemiological data and complex mathematical models. Therefore, epidemiological forecasting serves as an interface between quantitative disease dynamics and practical decision-making, translating data and uncertainty into information that can support informed actions [1].

Seasonal influenza serves as a well-established setting for studying epidemiological forecasting. Since the 2013/2014 season, the U.S. Centers for Disease Control and Prevention (CDC) has coordinated the FluSight forecasting challenge, in which participating teams submit probabilistic forecasts for influenza-related targets on a weekly basis. For example, teams provide forecasts of influenza-like illness (ILI) prevalence one to four weeks ahead. These forecasts are evaluated retrospectively against observed data and have contributed to the development of standardized forecasting targets and evaluation practices in epidemiology [2]. While the COVID-19 pandemic further highlighted the importance of forecasting for decision-making, influenza forecasting differs fundamentally from early COVID-19 forecasting, as it benefits from long historical time series and relatively stable seasonal dynamics, which enable more stable forecasting models.

Influenza activity is commonly monitored using surveillance indicators such as influenza-like illness, defined as the proportion of primary healthcare visits involving fever and respiratory symptoms. Although ILI is not specific to influenza and may therefore reflect illness caused by other respiratory pathogens, it remains a widely used and consistently reported indicator for influenza that enables comparisons across models, seasons, and locations [2]. In the FluSight initiative, forecasts are formulated explicitly for ILI.

In epidemiological forecasting, forecasts are most frequently used at short-term horizons. Short-term forecasts are defined as predictions one to four weeks ahead. These horizons are particularly relevant for near-term decisions such as hospital preparedness and resource planning and are generally more reliable than long-term projec-

tions, which are subject to greater uncertainty and structural changes due to evolving disease dynamics, as observed during the COVID-19 pandemic [3].

Modern epidemiological forecasting has increasingly emphasized probabilistic forecasts rather than point forecasts. Rather than predicting a single future outcome, probabilistic forecasts assign probabilities to a range of possible outcomes, thereby representing forecast uncertainty. To further improve predictive performance, ensemble forecasting approaches have been implemented, combining diverse modeling assumptions and data sources [2, 4].

To assess epidemiological forecasts, a variety of statistical evaluation metrics and scoring rules are used. These metrics compare forecasts to observed outcomes and thereby define a loss as the discrepancy between prediction and observation. Commonly used metrics include the logarithmic score, the continuous ranked probability score (CRPS), and prediction interval coverage. Historically, forecasts submitted to the CDC FluSight initiative were evaluated using a modified logarithmic score, which formed the basis for ranking forecasting models [2]. In practice, forecasts in the FluSight initiative are submitted in the form of predictive quantiles rather than full probability distributions. This makes the weighted interval score (WIS) particularly relevant, as it provides a proper scoring rule that is equivalent to CRPS under quantile-based reporting and has therefore become a standard metric in epidemiological forecasting [3].

Traditional evaluation metrics define expense implicitly through the chosen scoring rule and focus on discrepancies between forecasts and observations. While this approach provides valuable information about forecast quality, it does not explicitly account for how forecasts are used in decision-making or for the consequences of acting on them.

This observation has long been recognized in meteorology, where forecasts are routinely used to support decisions under uncertain conditions. In weather forecasting, cost-loss frameworks have been developed to evaluate forecasts in situations where decision-makers must balance the cost of preventive action against the potential loss incurred if no action is taken and an adverse event occurs [5, 6]. Within this framework, the usefulness of a forecast depends not only on its statistical accuracy but also on the decisions it supports and the consequences associated with those decisions.

Furthermore, this has motivated recent work on context-sensitive and decision-based evaluation frameworks. In particular, [7] implemented the cost-loss framework to evaluate infectious disease forecasts in influenza. The study considered the problem of predicting whether the seasonal influenza peak would exceed a predefined severity threshold, which means converting a probabilistic forecast into a binary decision problem.

Using this framework [7] introduced a Value Score that quantifies the usefulness of

a forecast within a specific decision setting relative to a historical baseline strategy and perfect information oracle. Their results showed that model rankings based on forecast value differed from rankings obtained using the modified log score employed in the FluSight challenge. This observation serves as the primary motivation for this thesis, which investigates the relationship between forecast accuracy and the value of forecast in short-term influenza forecasting.

1.1 Aim and Research Questions

This thesis aims to investigate the relationship between forecast accuracy and the value of forecast in short-term influenza forecasting. By comparing model performance under traditional forecast accuracy metrics and the value of forecast framework based on cost–loss theory, the study seeks to determine whether models that perform well according to traditional forecast accuracy metrics also perform well under a decision-sensitive value of forecast framework. More specifically, the thesis addresses the following research questions:

1. How do forecasting models perform according to traditional forecast accuracy metrics?
2. How do forecasting models perform according to the Value of Forecast framework?
3. To what extent do the resulting model rankings agree?

2

Theory

2.1 Epidemiological Forecasting

Epidemiological forecasting is an area concerned with predicting the progression of infectious diseases based on past and current observations, using a range of different methods. These methods are constructed using different approaches, including statistical methods based on historical observations, similar to those used in financial time series analysis, as well as mechanistic methods based on the underlying dynamics of disease transmission [2]. A commonly used mechanistic framework in epidemiological forecasting is compartmental models. These models divide the population into groups that are denoted compartments and models the flow between the compartments. For example, the SIR model divides the population into susceptible, infected, and recovered compartments [2]. In addition, ensemble methods have become increasingly popular, combining multiple models into a single forecast, often through different weighting schemes [2].

Disease prevalence can be measured using various metrics, one common example being ILI, which is an acronym for influenza-like illness (ILI). ILI is defined as a non-specific respiratory infection characterized by symptoms such as fever, cough, sore throat, fatigue, and muscle aches [8]. It is one of the most frequently used indicators of influenza activity in epidemiological surveillance and is measured as the weighted percentage of outpatient visits presenting influenza-like symptoms [9].

Other metrics may also be used, such as hospitalizations and virologic testing [2]. Depending on the application, different aspects of disease prevalence may be of interest, for example, forecasting ILI one month ahead or perhaps the timing of the seasonal peak [2]

As a result, both point forecasts and probabilistic forecasts are relevant in epidemiological forecasting. Point forecasts can be useful when a single predicted outcome is of primary interest, such as predicting the most likely peak week of an epidemic season. However, for short-term forecasts, such as predicting ILI four weeks ahead, there is often substantial uncertainty. For decision-makers, an exact predicted value is therefore less informative than a probabilistic forecast, which provides a distribution over possible outcomes and highlights regions of higher likelihood [2, 3].

Consequently, epidemiological forecasting plays an important role in healthcare plan-

ning, as forecasts can be used to anticipate short-term trends in disease activity and support decision-making regarding, for example, resource allocation [10]. However, traditional forecast evaluation metrics commonly focus on predictive accuracy. While such metrics provide information about how well forecasts match observed outcomes, they do not account for how forecasts are used in decision-making or whether improvements in forecast accuracy necessarily lead to better decisions. This motivates the need for frameworks that explicitly evaluate the value of forecasts in relation to decisions.

2.2 Probabilistic forecasts and the FluSight challenge

The Centers for Disease Control and Prevention (CDC) initiated the FluSight forecasting challenge as a collaboration between multiple research groups aimed at improving the quality of predicting the progression of influenza seasons across the United States. The objective of the challenge was to provide a platform where groups could submit weekly forecasts, which are then aggregated and evaluated for influenza-like illness (ILI), both at the regional and national levels. The challenge was introduced to support public health decision-making by improving situational awareness of influenza activity [9].

Within the FluSight challenge, groups provide multiple types of forecasts each week. These include short-term horizon forecasts, which focus on predicting weekly ILI values 1 to 4 weeks ahead, as well as seasonal targets such as the onset week, the peak ILI percentage, and the timing of the seasonal peak. It has been consistently observed that ensemble models, which combine multiple individual forecasts, often outperform single-model approaches [9].

Most importantly, short-term horizon forecasts are expressed probabilistically, as distributions over possible ILI values rather than as point predictions. This probabilistic representation captures the uncertainty in disease progression and provides a basis for decision-making under uncertainty, which is central to this thesis [9].

To illustrate the concept of probabilistic short-term forecasts within the FluSight framework, we use Figure 2.1 as a synthetic example. At a given forecasting week, observations are available only up to the current time point. Based on this information, a forecasting model produces predictive distributions for future ILI values at different forecast horizons. In Figure 2.1, forecasts are generated one and four weeks ahead. The predictive distributions represent the uncertainty in future influenza activity, whereas the true ILI values only become available once the weeks have been observed. Subsequently, forecast performance can be determined by comparing the observed outcomes with the forecast distribution.

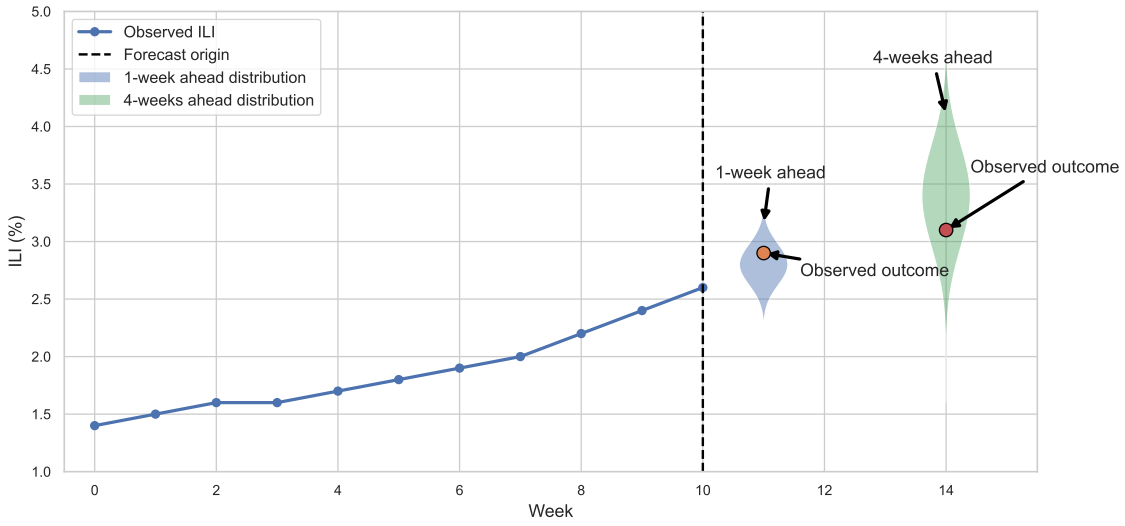

Figure 2.1: Illustrative example of probabilistic ILI forecasting. Historical observations are available up to the forecast origin, after which predictive distributions are generated 1 and 4 weeks ahead. The realized ILI values are observed only after the target weeks occur and can then be compared with the forecast distributions.

ILI values are expressed as percentages, representing the proportion of healthcare visits attributed to influenza-like illness. For example, an ILI value of 1 corresponds to 1%. During off-season periods, ILI values are typically low, below 2%, and relatively stable, whereas higher values are observed during the influenza season when activity increases.

In FluSight, these distributions are represented using discretized bins of width 0.1 % over the ILI range, where each bin is assigned a probability. The range of ILI is partitioned from 0 % up to 13 %, with an additional bin covering values from 13 % to 100 % to capture the more extreme outcomes [9, 2].

2.3 Forecast Accuracy Evaluation

To evaluate the accuracy of a forecast, there exist several metrics used to measure how well a forecast agrees with observed outcomes. Both point and probabilistic forecasts can be evaluated. One of the most common evaluation metrics for point forecasts is the Mean Absolute Error (MAE), defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{2.1}$$

where $\hat{y}_i$ denotes a model prediction and y_i the observed outcomes. The score is averaged over all observation-prediction pairs. A lower MAE, therefore, corresponds to a better forecast [11].

For probabilistic forecasts, scoring rules are often used in place of point-error metrics. A proper scoring rule is defined as a scoring rule where the expected score is maximized when the forecaster reports their true predictive distribution. Therefore, a proper scoring rule doesn't incentivize a forecaster to report forecasts that differ from their beliefs [4].

In the FluSight challenge, a commonly used metric is the log score, which rewards models assigning a high probability to the event that is subsequently observed. It is defined as

$$\text{LS} = \frac{1}{n}\sum_{i=1}^{n} \log P(X_i \in B(y_i)). \tag{2.2}$$

Here, X_i denotes the forecasted random variable for the forecast instance, and $B(y_i)$ denotes the bin (or set of bins) containing the observed value y_i. Therefore, a higher log score corresponds to forecasts assigning greater probability mass close to the observed outcome. As with MAE, the score is summed over all observation-prediction pairs [2].

Furthermore, the log score has also been used in an extended form. The score above is sometimes referred to as the single-bin log score; one could also extend this to a multi-bin log score. In this case, $B(y_i)$ is extended to include adjacent bins often symmetrically around the observed bin. For example, one can include three bins above and three bins below the observed outcome. While the single-bin log score is proper, the multi-bin log score is not proper [2].

Another popular scoring rule is the weighted interval score. It is a proper scoring rule based on a weighted sum of prediction intervals. Each interval penalizes overprediction, underprediction, and forecast dispersion, and it is defined as

$$\text{WIS}(F, y) = \frac{1}{K + \frac{1}{2}} \left(\frac{1}{2}|y - m| + \sum_{k=1}^{K} \frac{\alpha_k}{2} \cdot \text{IS}_{\alpha_k}(F, y) \right) \tag{2.3}$$

where the interval score at significance level α is

$$IS_\alpha(F, y) = (u - l) + \frac{2}{\alpha}(l - y)\mathbf{1}(y < l) + \frac{2}{\alpha}(y - u)\mathbf{1}(y > u). \tag{2.4}$$

Here, y is the observed value, m is the predictive median, l and u are the lower and upper bounds of the prediction interval, α is the significance level, and K is the number of intervals used. Therefore, a lower WIS corresponds to a better forecast [3].

However, forecast accuracy may not be telling the entire story. Even though a model produces a high forecast accuracy over many observations, it may fail in some instances. Precisely when the model fails may correspond to the most important scenarios for a stakeholder using the forecast. Therefore, an event-based forecasting measure that is decision-context sensitive may provide additional information. We now proceed by defining event scenarios.

2.4 Event scenarios

In the FluSight challenge, each model provides a probabilistic forecast over bins. As mentioned in Section 2.2, these forecasts are relatively precise with bins of size 0.1 %. However, a decision maker is often not concerned with the exact forecasted values, but rather with whether a certain event will happen. For instance, one may be interested in whether the influenza season will exceed a given threshold, which could motivate additional allocation of perhaps ICU resources or other hospital capacity measures to support a more vulnerable hospital [10].

In this thesis, such events are defined in a relative manner. Let Y_t denote the current week's ILI and Y_{t+h} denote the future ILI at time $t+h$, where h denotes the forecast horizon. Furthermore, let γ be a relative growth parameter. We are then concerned with whether or not this condition is true or false,

$$Y_{t+h} \geq (1+\gamma)Y_t. \tag{2.5}$$

Given a predictive distribution for Y_{t+h}, each model can therefore assign a probability to the event in (2.5). (2.5) thus converts a continuous forecasting problem into a binary event prediction problem and can be interpreted as whether the future ILI level increases by at least γ relative to the current ILI. A low value of γ corresponds to an event with a small increase in ILI, whereas a high value of γ corresponds to an event with a more substantial increase in ILI. Hence, different levels of γ represent different levels of concern regarding future increases in influenza activity.

2.5 Cost–loss scenario

Consider an administrator at a hospital who is in charge of allocating staff and resources. The administrator receives short-term forecasts indicating an elevated probability of a substantial increase in influenza activity during the coming weeks. Based on this scenario, the administrator is tasked with deciding whether to take preventive action, which may include more staff, additional ICU beds or simply reallocating medical resources. These measures incur a cost even if the forecasted event doesn't occur. Conversely, if no preventive action is taken and the influenza activity rises substantially, the hospital may face shortages of staff, medical resources, and other critical resources. The decision therefore involves balancing the cost of preparedness against the potential consequences of being unprepared.

This trade-off is formalized through the cost-loss framework [5, 6]. Since we have defined an event, it is now up to the stakeholder to decide whether an action should be taken. This decision is based on the forecast that the model produces regarding whether the event will occur or not, however we need some sort of rule that helps us decide when to act or not. Therefore, the forecast is not determined solely by the forecast but the stakeholder must weigh the cost of acting preventively against the potential loss incurred if no action is taken and the event occurs [5, 6]. Thus, in this

thesis, we could consider a binary world where acting means taking preventive action resulting in a cost denoted as C and not acting resulting in no cost. However, these actions are then measured against a loss, incurred if no preventive action is taken and the event occurs, which is denoted as L. One can also take preventive action where the event doesn't occur, therefore resulting in a cost of C. The resulting cost-loss matrix is shown in Table 2.1.

Table 2.1: Cost–loss matrix for a binary decision problem. C denotes the cost of preventive action and L the loss incurred if no action is taken and the event occurs.

	Event occurs	Event does not occur
Take action	C	C
Do not take action	L	0

We will assume that $0 < C < L$, and with this assumption, we can deduce that the rule determining taking preventive action will be fulfilled if our forecast $0 < p < 1$ fulfills $C > pL$ which can be rearranged into $p > \frac{C}{L}$. Therefore, if our decision maker argues that the ratio between cost and loss is low, the threshold for taking preventive action is low. Compared to a high cost-loss ratio, the forecast must assign a lot of probability to the event to fulfill the threshold.

To illustrate this, consider a situation where the cost of preventive action is $C = 1$ and the loss is $L = 10$, resulting in a cost–loss ratio $r = \frac{C}{L} = 0.1$. If a forecast assigns a probability $p = 0.3$ to the event, then $p > r$, and action is taken. Conversely, if the forecast probability is $p = 0.05$, then $p < r$, and no action is taken. This example highlights how the decision depends jointly on the forecast probability and the cost–loss ratio.

Therefore, the usefulness of a model and its forecast not only depends on the predictive accuracy but also on the stakeholders' selected cost-loss ratio. Now, we have weighed our decision against the forecasted probability and the cost of acting, but now we need to understand how much benefit this forecast may provide relative to a baseline that has no access to a model forecast [5, 6].

2.6 Value of Forecast

While Section 2.5 describes how decisions can be made using forecasted probabilities, it doesn't quantify how valuable a forecast is relative to a baseline. Thus, a model forecast provides value if it can improve the decisions being made compared to a baseline that relies solely on historical data and not on having access to any forecasts.

A model that often provides accurate forecasts may not always be beneficial, since the decision context matters. If an otherwise accurate model fails during rare but crucial events, this may be very costly. Meanwhile, a slightly less accurate model

may be better at predicting decision-relevant events.

We therefore need a means to quantify the value of a forecast. A natural idea is thus to compare decision outcomes with different access to information. Firstly, decisions are solely based on historical information. Secondly, the decisions are based on a model forecasting probabilities. Thirdly, decisions are based on an idealized decision maker that already knows the future outcome of an event and thus always takes the optimal action. For each of these situations, it's possible to compute an expected expense with regard to the cost-loss framework [5, 6].

A forecast provides improvement relative to a baseline if the expected expense under forecast-based decisions is lower than under the baseline strategy. To facilitate interpretation, this improvement is normalized by the improvement achieved under perfect foresight relative to the same baseline. This leads us naturally to defining the value of a forecast as

$$\mathrm{VS} = \frac{E_b - E_f}{E_b - E_p}. \tag{2.6}$$

where E_b denotes the expected expense under a baseline that doesn't have access to forecasts. It is often called a climatological benchmark in meteorological forecasting. Furthermore, E_f denotes the expected expense when decisions are made using a forecast, and E_p is the expected expense under an oracle, a decision maker who always knows the correct action. Hence, the numerator measures the gain of using a forecast relative to a baseline, while the denominator measures the biggest possible gain, therefore creating a normalized measure [5, 6].

A $\mathrm{VS} = 1$ is equivalent to a perfect forecast, while $\mathrm{VS} \approx 0$ indicates a model equivalent to the baseline. lastly, $\mathrm{VS} < 0$ indicates a model worse than the baseline. The range of VS is thus given by $(-\infty, 1]$. However, the value of a forecast depends on the cost-loss ratio. Thus, a model may, for certain cost-loss ratios, take on positive values of VS, while for other cost-loss-ratios it may produce negative VS [5, 6].

A forecast that achieves high forecast accuracy may still have limited value if it doesn't improve decisions relative to the baseline. On the contrary, a forecast with moderate accuracy may generate greater value if it can better support decisions relative to all events. Comparing models with respect to forecast accuracy and value of forecast is therefore of interest, which motivates us to introduce a tool that can measure the similarity or difference between rankings of models created by the two approaches.

2.7 Rank correlation

In this thesis, we are interested in comparing forecast accuracy, such as MAE, LS, and WIS, to VS. Each of these approaches induces a ranking of the forecasting models, picking up different aspects of the forecasts. This motivates us to quantify

how similar and dissimilar the resulting rankings are.

To measure the agreement between two ranking systems, there are multiple ranking correlation coefficients that can be used. In this thesis, Spearman's rank correlation coefficient is applied. The coefficient is similar to the standard correlation coefficient since its value range is $[-1, 1]$ and it measures the monotonic association between two rankings and is defined as

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}, \tag{2.7}$$

where d_i denotes the rank difference for model i, and n is the total number of models being compared [12].

A value of $\rho = 1$ corresponds to identical rankings between the two approaches. A value of $\rho = 0$ indicates no monotonic agreement, meaning that the relationship between the ranking systems is weak or more or less random. A value of $\rho = -1$ corresponds to perfectly reversed rankings [12].

Therefore, a low rank correlation would indicate that evaluating models solely through traditional forecast accuracy metrics may overlook models that provide greater decision value.

3

Methods

3.1 FluSight Data Repository

As outlined in Section 2.2, this thesis is concerned with data that originates from the CDC FluSight challenge, a collaborative forecasting initiative that involves researchers and public health officials, which is aimed at producing real-time forecasts of influenza activity in the United States. The empirical data used in this thesis were obtained from a publicly available GitHub repository [13], which contains historical forecast submissions from the season 2010/2011 up to the 2018/2019 season with contributions from more than 20 forecasting models.

However, even though a historical repository serves as the basis of this thesis, the FluSight initiative is still an ongoing project. For example, there exists a repository that collects forecasts for the 2025-2026 season [14]. However, these newer systems are based on different surveillance targets, such as laboratory-confirmed influenza tests, rather than the ILI metric considered in this thesis.

The repository [13] is organized around forecast submissions from individual models, including both ensemble models and component models. For each model, forecast submissions are stored in comma-separated value (CSV) format. Here, each file corresponds to one model's forecast for a specific epidemiological week (EW) in a given year. An influenza season runs from EW40 of a year till EW20 of the following year. Files are named according to the structure "EWxx-yyyy-model.csv", where "xx" denotes the epidemiological week, "yyyy" denotes the year, and "model" is the forecasting model. Each forecast file contains information regarding the week the forecast was issued, forecast horizon, the geographic target, and forecast type, which are provided either as point predictions or as probabilistic forecasts represented through bins with assigned probability mass. In addition, a forecast may be a short-term forecast, such as a one to four week ahead forecast, or a seasonal forecast, such as the season's onset week, peak percentage, or peak week.

Furthermore, the repository [13] contains a file named "target-multivals.csv", which stores observed ILI values, used as ground truth outcomes. This file also follows a structure that is compatible with the forecast submissions and is used both to evaluate forecast accuracy and construct the event-based decision framework introduced in Section 2.4.

The analysis will be restricted to short-term forecasts made one to four weeks ahead for the region ”US National’. Included seasons depend on forecast availability, since some models do not provide submissions for all seasons.

3.2 Model Selection

The historical FluSight repository contains submissions from a large number of forecasting models across multiple influenza seasons. Rather than analyzing all available models in this thesis, a subset of models was selected in order to obtain a focused and comparable evaluation.

The primary reasoning behind the model selection was to remain consistent with the model set considered in a related study [7], which, similarly to this thesis, investigates the value of influenza forecasting models in a cost-loss setting.

The selected models consist of one ensemble model and eight individual component models. Together, these models represent a range of forecasting methodologies that were used within the FluSight challenge, including statistical learning methods, density estimation approaches, Bayesian averaging methods, and mechanistic compartmental models.

Model names often reflect the submitting research group or institution, followed by the specific forecasting methodology. The model set considered in this thesis is presented in Table 3.1, alongside a short description of each model, which was obtained from [15].

Model	**Main methodology**
CUBMA	Bayesian Model Averaging
ReichLab_kde	Kernel Density Estimation with penalized spline
Delphi_ExtendedDeltaDensity	Delta Density model
Delphi_MarkovianDeltaDensity	Markovian Delta Density model
LANL_DBMplus	Dynamic Bayesian SIR model with discrepancy
ReichLab_kcde_backfill_post_hoc	Kernel Conditional Density Estimation
ReichLab_kcde_backfill_none	Kernel Conditional Density Estimation
CU_EAKFC_SIRS	Ensemble Adjustment Kalman Filter SIRS model
target-type-based-weights	Ensemble model using target-specific weights

Table 3.1: Forecasting models included in the analysis.

3.3 Data Preparation

As mentioned in Section 3.1, the forecast data are provided as CSV files containing submissions for each epidemiological week, year, and model. In addition, observed ILI values serving as ground truth are available in a similarly structured format. The objective of this Section is therefore to aggregate the information into a format that is suitable for analysis.

The first step is to filter the data according to the scope of this thesis. As previously described in Section 3.1, this thesis focuses on the forecasting models defined in Section 3.2 that are short-term forecasts for the region "US National'. Therefore, forecasts from other regions as well as seasonal targets are excluded. The remaining data are then split into two datasets: one containing point forecasts and one containing probabilistic forecasts represented through bins. Thereafter, all forecast submissions are aggregated into unified datasets, where each row corresponds to a unique forecast instance. Each forecast instance is represented by a combination of epidemiological week, year, season, model, and forecast horizon alongside its prediction, either a point prediction or a bin with an associated probability mass.

To evaluate the forecast, both datasets are merged with the ground truth, that is, the observed ILI values. The merge is performed using epidemiological week, year, and horizon, which ensures that each unique forecast is aligned with its realized outcome.

While this structure is sufficient for evaluating the forecast accuracy, the event-based value of forecast framework requires an additional operation. We must transform the ground truth data, since the data are stored in a format that corresponds to future outcomes. However, the event definition introduced in Section 2.4 requires us to compare the current value Y_t with a future value Y_{t+h}.

Therefore, a new dataset is constructed where each observation Y_t at time t is paired with its corresponding future value Y_{t+h} at time $t+h$ for each forecast horizon. This allows us to determine whether the event occurs and enables a comparison between forecasts and their realized outcomes within the event-based value of the forecast framework.

3.4 Forecast Accuracy Evaluation

The next step is to evaluate the forecasts using the prepared datasets from Section 3.3, which are the point forecast dataset and the probabilistic forecast dataset.

The mean absolute error (MAE) is used to evaluate point forecasts. The MAE is computed over all prediction-observation pairs for each model, (2.1). The results are then aggregated separately for each forecast horizon.

For probabilistic forecasts, we use the discretized bin-based representation provided

by the FluSight data. The log score (LS) is computed using the single-bin approach, where the probability mass assigned to the bin containing the observed ILI value is evaluated, (2.2). Similar to MAE, the results are aggregated separately per horizon.

The weighted interval score (WIS) is also computed for probabilistic forecasts. The score is computed based on a set of significance levels, defined as $\alpha = [0.02, 0.05, 0.1, \ldots, 0.95]$ from which each significance level gives rise to a prediction interval, (2.3). As with the other metrics, the results are aggregated separately per horizon

In total, three forecast accuracy metrics are applied: one point-based (MAE) and two probabilistic metrics (LS and WIS). Lower values of MAE and WIS correspond to better forecasts, whereas a higher LS indicates a better forecast.

3.5 Event Construction

As introduced in Section 2.4, an event is said to have occurred if

$$Y_{t+h} \geq (1 + \gamma)Y_t. \tag{3.1}$$

where Y_t denotes the current week's ILI % and Y_{t+h} denotes the ILI % value h weeks ahead.

As mentioned in Section 3.3, the original observed data is structured in terms of future outcomes corresponding to its forecast targets. However, the event definition, as seen in (3.1), requires a comparison between the current observation Y_t and a future observation Y_{t+h}. Therefore, the ground truth data must be transformed into a format where we can form such pairs. To achieve this, we create a time index within each season referred to as *season time*. This index orders epidemiological weeks sequentially within each season. With this, we can align observations across different forecast horizons. In doing so, special care is taken when forecasts wrap across years and when years contain 53 weeks.

Using this representation, a dataset is constructed where each observation at time t is paired with its corresponding future values at time $t + h$, for each $h = 1, 2, 3, 4$. This makes it possible to determine whether (3.1) occurs, thereby defining a binary variable. In addition to determining whether an event occurs, we also compute the forecasted probability of an event, p_f, which is obtained by summing the probability mass assigned to all bins exceeding the threshold set by $(1 + \gamma)Y_t$.

The parameter γ controls the severity of an event. In the thesis, we consider $\gamma \in [0, 0.5]$, which spans from small fluctuations around $\gamma = 0$ up to more substantial increases such as $\gamma \approx 0.5$. That is, γ represents the minimum relative increase in ILI h weeks ahead required for an event to occur.
The event definition (3.1) can equivalently be rewritten as $\frac{Y_{t+h}}{Y_t} \geq (1 + \gamma)$. This interpretation is useful because the quantity $\frac{Y_{t+h}}{Y_t}$ represents the observed growth rate

between the current and future ILI level. Consequently, the empirical distribution of growth rates provides information about how common or rare different event thresholds are in the observed data.

3.6 Cost–loss implementation

Now we have information on whether an event occurred and also information about how probable our forecast considers the event to be. This allows us to make a decision for each forecast instance. As introduced in Section 2.5, Table 2.1 describes the possible outcomes. The choice of action is dependent on the forecast probability, p_f, and the predetermined cost-loss ratio, r. That is, we take action if $p_f > r = \frac{C}{L}$. Strategy-wise, a low r corresponds to a belief that actions are relatively cheap, leading to more frequent interventions, while a larger r corresponds to a belief where preventive action is more expensive and thus action is taken less often. That is, for each forecast instance (t, h), a decision is made based on whether the forecasted probability p_f exceeds the threshold r.

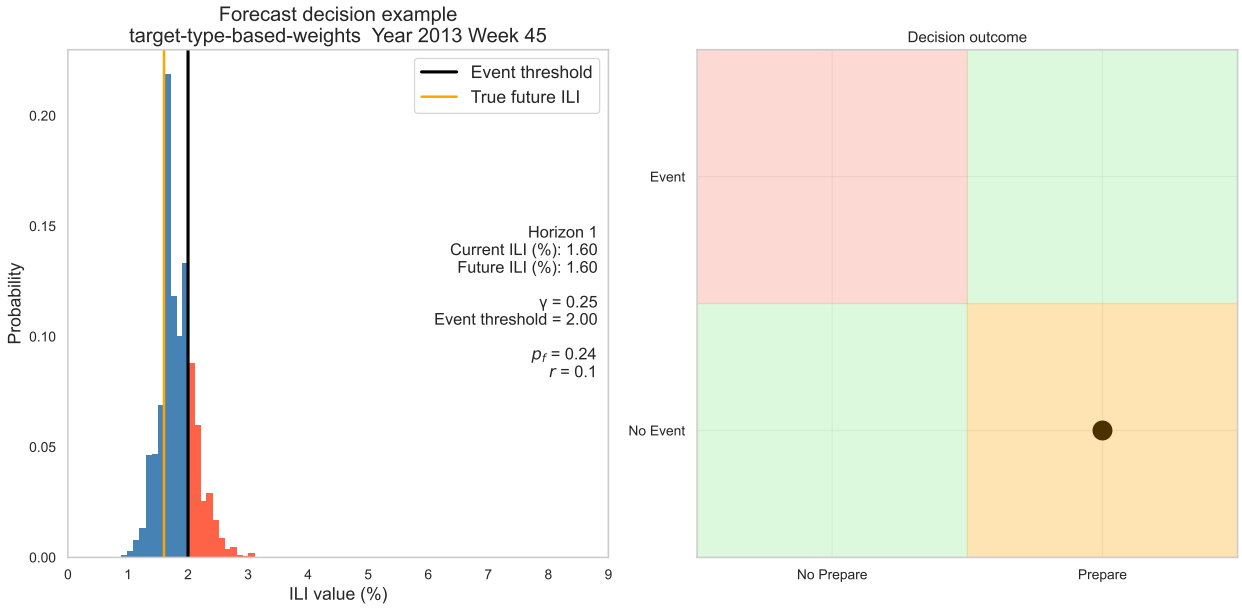

Figure 3.1: Example of the cost–loss decision framework for a single forecast instance with event definition $\gamma = 0.25$, forecast horizon $h = 1$, and cost–loss ratio $r = 0.10$. The left panel shows the probabilistic forecast together with the event threshold used to compute the forecast probability p_f, while the right panel displays the corresponding decision outcome in the cost–loss matrix.

To illustrate how this decision rule is applied in practice, Figure 3.1 shows an example based on a single forecast instance. The left panel displays the probabilistic forecast together with the event threshold, where the forecast probability p_f is computed as the probability mass above the threshold. This probability is then compared to the cost–loss ratio r to determine whether action is taken.

The right panel shows the corresponding outcome in the cost–loss matrix. Depending on whether the event occurs and whether action is taken, a cost is assigned according to Table 2.1. In the illustrated example, action is taken even though the event does not occur, resulting in a false alarm and an incurred cost of C.

Hence, the illustrated example corresponds to one of the four possible outcomes in the cost–loss framework. For completeness, examples of the remaining decision outcomes are provided in Appendix B.

Furthermore, in section 2.6, three types of expenses were introduced based on different levels of information. These are the forecast-based expense, the baseline expense, and the perfect foresight expense. The forecast-based expense, E_f, is computed by applying the decision rule to each forecast instance and assigning a corresponding cost or loss according to table 2.1. The expected expense is then obtained by averaging across all forecast instances.

The baseline expense, E_b, is computed using a historical estimate of the event probability denoted as p_b, instead of the model-based probability p_f. This corresponds to decisions made without access to model-based forecasts.

The perfect foresight expense, E_p, represents the minimum achievable expense and corresponds to a decision-maker who only acts when the event occurs.

Together, these three expenses provide the basis for evaluating the decision value of forecasts in the following section.

3.7 Value of Forecast Computation

Now we compute the value of forecast using our previously defined expenses: the forecast-based expense, E_f, the baseline expense, E_b, and the perfect foresight expense, E_p. The computation is carried out across different event definitions $\gamma \in [0, 0.5]$, cost–loss ratios $r \in [0.05, 0.95]$, forecasting models, and forecast horizons.

We justify choosing the range $r \in [0.05, 0.95]$ to exclude extreme decision scenarios. When $r \to 0$, the cost of taking action is negligible relative to the loss, which leads to a strategy where action is always taken. Conversely, when $r \to 1$, the cost of acting approaches the loss, making the decision effectively indifferent. In both cases, the decision problem becomes trivial and does not provide meaningful insight into the value of forecasts. Therefore, we restrict r to the interval $[0.05, 0.95]$ to ensure that we focus on non-trivial and practically relevant decision settings.

The baseline probability is a crucial component of the value of forecast framework. To obtain a realistic and meaningful baseline, we construct it using historical event

frequencies. Specifically, for each forecast horizon, the event frequency is first computed within each season and then aggregated over past seasons using an expanding mean. That is, only observations from previous seasons are used when computing the baseline expense. This results in a time-varying baseline probability p_b that reflects a changing historical likelihood of the event.

For each model, forecast horizon, event definition, and cost-loss ratio, the decision rule is applied using both the forecast probability p_f and the baseline probability p_b. The corresponding expenses E_f, E_b, and E_p are computed for each forecast instance and then averaged over all instances.

More precisely, let N denote the total number of forecast instances, and let $E^{(i)} \in \{0, 1\}$ denote whether the event occurs for instance i. Then the forecast-based expense is given by

$$E_f = \frac{1}{N} \sum_{i=1}^{N} \begin{cases} C, & \text{if } p_f^{(i)} > r \\ L, & \text{if } p_f^{(i)} \leq r \text{ and } E^{(i)} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

Similarly, the baseline expense is computed using the historical probability p_b as

$$E_b = \frac{1}{N} \sum_{i=1}^{N} \begin{cases} C, & \text{if } p_b^{(i)} > r \\ L, & \text{if } p_b^{(i)} \leq r \text{ and } E^{(i)} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

The perfect foresight expense corresponds to the expected loss under optimal decisions and is computed as

$$E_p = \frac{1}{N} \sum_{i=1}^{N} C \cdot E^{(i)}. \tag{3.4}$$

Based on these expenses, the Value Score is computed as

$$\text{VS} = \frac{E_b - E_f}{E_b - E_p}. \tag{3.5}$$

The Value Score is computed over all parameter combinations. That is, over all event definitions, γ, models, m, cost loss ratios, r, and forecast horizons h, which results in a stored dataset of value scores $\text{VS}(\gamma, m, r, h)$. The full procedure is summarized in Algorithm 1.

Algorithm 1 Computation of Value of Forecast

Require: Forecast distributions, observed data, set of thresholds γ, cost–loss ratios r, horizons h

Ensure: Value scores $\mathrm{VS}(\gamma, r, h, m)$

for each γ **do**
  Compute event threshold $(1+\gamma)Y_t$
  Compute forecast probabilities p_f from bin forecasts
  Compute binary event outcomes E
  Compute historical baseline probabilities p_b
  for each model and forecast horizon **do**
    for each cost–loss ratio r **do**
      Compute forecast-based loss E_f
      Compute baseline loss E_b
      Compute perfect foresight loss E_p
      Compute value score:

$$\mathrm{VS} = \frac{E_b - E_f}{E_b - E_p}$$

    end for
  end for
end for

3.8 Integrated Value Measures

We have now defined the value score $\mathrm{VS}(\gamma, r, h, m)$, which depends on the event definition γ, the cost–loss ratios r, the forecast horizon h, and the forecasting model m. Since the Value Score varies with the cost-loss ratio, evaluating model performance at a single value of r provides only a limited understanding of the model's usefulness, as it corresponds to a single decision setting.

To obtain a more concise and lower-dimensional representation of model performance, we integrate the value score across the range of cost-loss ratios. This is done numerically by approximating

$$IV(\gamma, h, m) = \int_r \mathrm{VS}(\gamma, r, h, m)\, dr, \tag{3.6}$$

using a discrete grid of $r \in [0.05, 0.95]$ together with the trapezoidal rule.

The integrated value $\mathrm{IV}(\gamma, h, m)$ summarizes how well a model performs across a range of decision scenarios rather than for a single cost-loss ratio. A higher value indicates that the model consistently yields better decisions than the baseline across diverse decision settings. Values close to zero indicate performance similar to the baseline, while negative values indicate that decisions based on the forecast perform worse than the baseline across different decision settings on average.

3.9 Rank Correlation Analysis

We have now computed both the forecast accuracy metrics as well as the Value Score, $\mathrm{VS}(\gamma, r, h)$ for each model. Each of these approaches induces a ranking of the forecasting models.

For the forecast accuracy metrics, rankings are obtained for each horizon h, where lower values of MAE and WIS indicate better performance, while higher values of LS indicate better performance. Similarly, models can be ranked according to the Value Score, where higher values correspond to better decision performance.

A key difference is that the forecast accuracy only depends on the forecast horizon h, whereas the Value Score depends both on the forecast horizon h, on the event definition γ, and the cost–loss ratio r. This results in a ranking of the Value Score based on each parameter combination.

To quantify the agreement between these ranking approaches, we compute the Spearman rank correlation between the forecast accuracy-based rankings and the Value Score rankings for each combination of γ, h, and r. This allows us to understand how consistently the different evaluation approaches rank the forecasting models, that is, how similar these ranking systems are to one another.

4

Results

In this chapter, we present the empirical results of the thesis. We begin by examining the characteristics of the influenza data and the event definitions, including seasonal patterns, growth rate distributions, and event frequencies. Next, the behavior of the baseline forecasts is investigated. We then evaluate the forecasting models using both traditional forecast accuracy metrics and the value of forecast framework. Finally, the relationship between forecast accuracy and forecast value is analyzed through rank correlation comparisons between the induced model rankings of the different metrics.

4.1 Event and Data Characteristics

4.1.1 Seasonal ILI trajectories

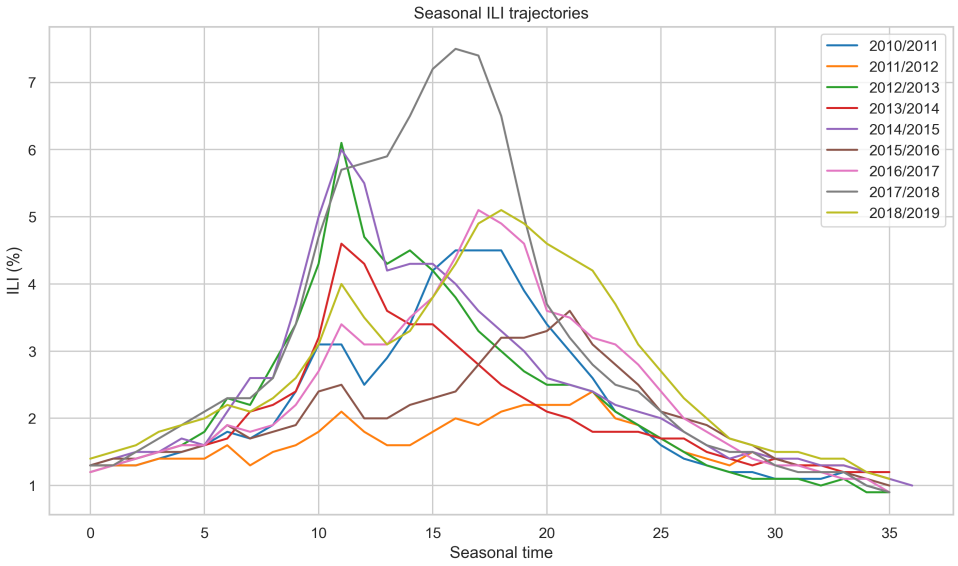

Figure 4.1: Observed ILI prevalence across influenza seasons as a function of *season time*. Each line corresponds to one influenza season.

In Figure 4.1, we can observe the observed ILI prevalence from the start of each season (*season time*=0) to the end of each season (*season time*≈35) for all seasons. Each line represents the ILI prevalence for a single season. Overall, we observe a similar seasonal structure: the beginning and end of the seasons show low ILI levels,

with varying peak magnitudes and peak timing, but overall, a peak during the winter period. Thus, while the seasons exhibit a common seasonal pattern, there is still variation between each season.

4.1.2 Growth rate distribution

Since an event occurs whenever $\frac{Y_{t+h}}{Y_t} \geq 1+\gamma$, the distribution of growth rates provides insight into how frequently different event definitions occur in the observed data. $\frac{Y_{t+h}}{Y_t}$ represents the growth rate between the current ILI value and the future ILI value at forecast horizon h. Using the observed ground truth data, we can therefore examine the empirical distribution of these growth rates.

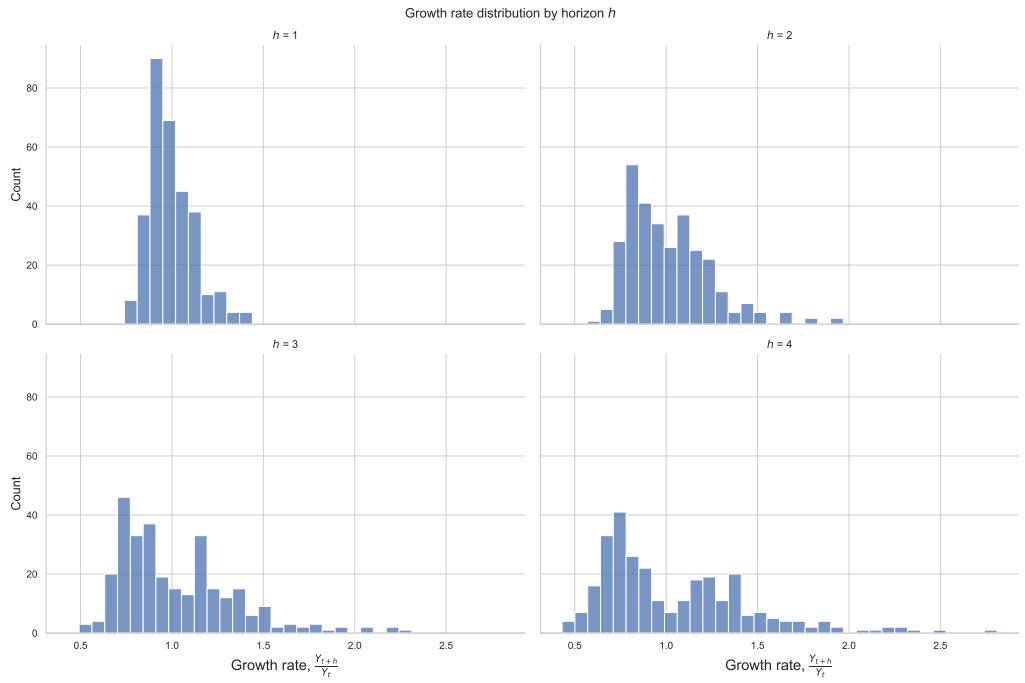

Figure 4.2: Empirical distribution of the growth rate $\frac{Y_{t+h}}{Y_t}$ across forecast horizons $h = 1, 2, 3, 4$.

From Figure 4.2 we observe that most growth rates are concentrated around 1, indicating that the ILI value at time $t+h$ is often similar to the current value at time t. However, the distributions are right-skewed, meaning that growth rates greater than 1 occur more frequently than large decreases. Thus, increases in ILI are more common than equally large decreases across the observed current and future week pairs.

Furthermore, the distributions become wider as the forecast horizon increases. This indicates that larger deviations from the current ILI value become more common as the forecast horizon increases.

4.1.3 Event frequency by horizon

We now examine how the event frequency varies with respect to the event definition parameter γ and the forecast horizon h. Here, the event frequency denotes the proportion of all current and future week pairs that satisfy (3.1) for a given γ and horizon h.

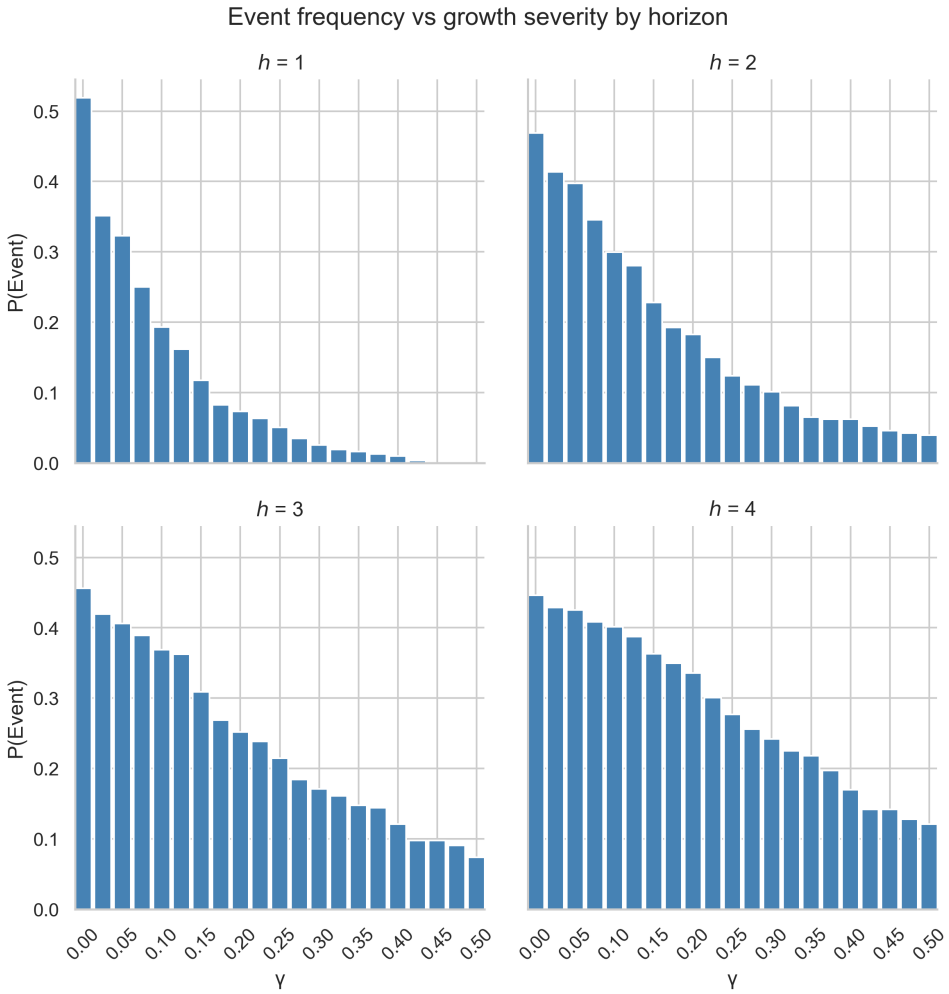

Figure 4.3: Event frequency as a function of the event definition parameter γ across forecast horizons $h = 1, 2, 3, 4$.

From Figure 4.3 we observe that the event frequency decreases as γ increases. Furthermore, the decrease in event frequency differs across forecast horizons. For longer horizons, larger growth events have a bigger event frequency compared to shorter horizons. Lastly, for smaller values of γ, the event frequencies remain relatively similar across all forecast horizons.

4.1.4 Event probability by season

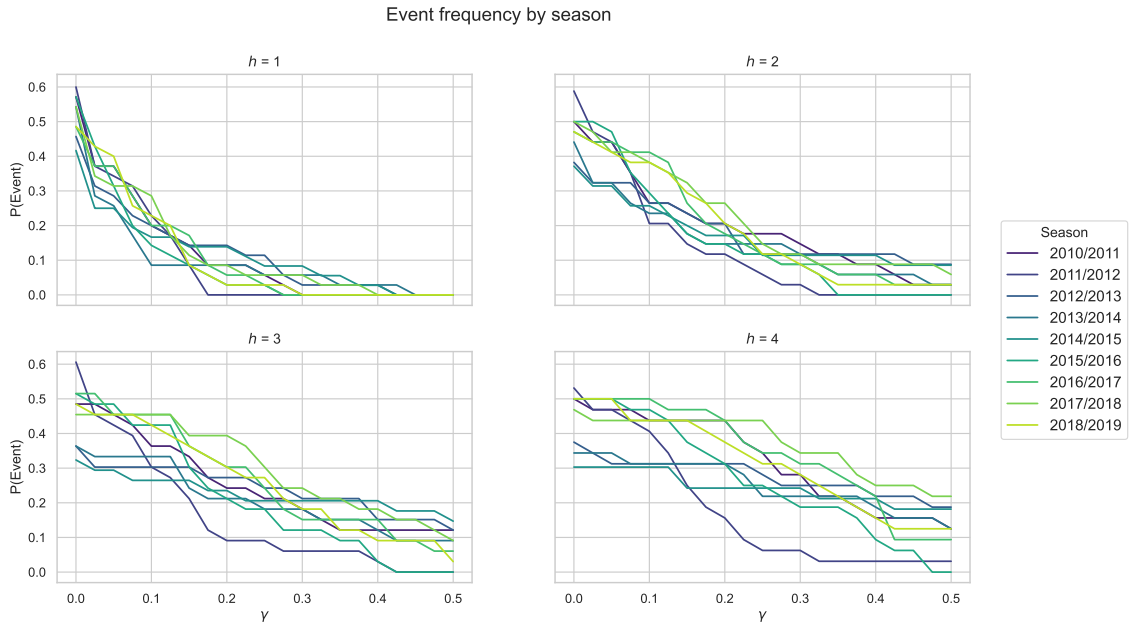

Figure 4.4: Empirical event frequencies across influenza seasons as a function of the event definition parameter γ for different forecast horizons.

Figure 4.4 shows the empirical event frequencies for all seasons across γ, where each line corresponds to one influenza season. From Figure 4.4 we observe that the seasons display relatively similar behavior overall, although some seasons, such as 2010/2011 for example, exhibit somewhat different patterns compared to the other seasons.

In general, the event frequency for each season decreases as γ increases. However, for larger forecast horizons, the decrease is less rapid compared to shorter horizons.

4.2 Baseline Behavior

Recall that the baseline represents a forecast based solely on historical event frequencies and does not incorporate any current epidemiological information. Two baseline variants are considered: a static baseline probability, which is computed using all available seasons, and then we have a historical baseline probability, which is computed using only information from previous seasons.

4.2.1 Historical belief

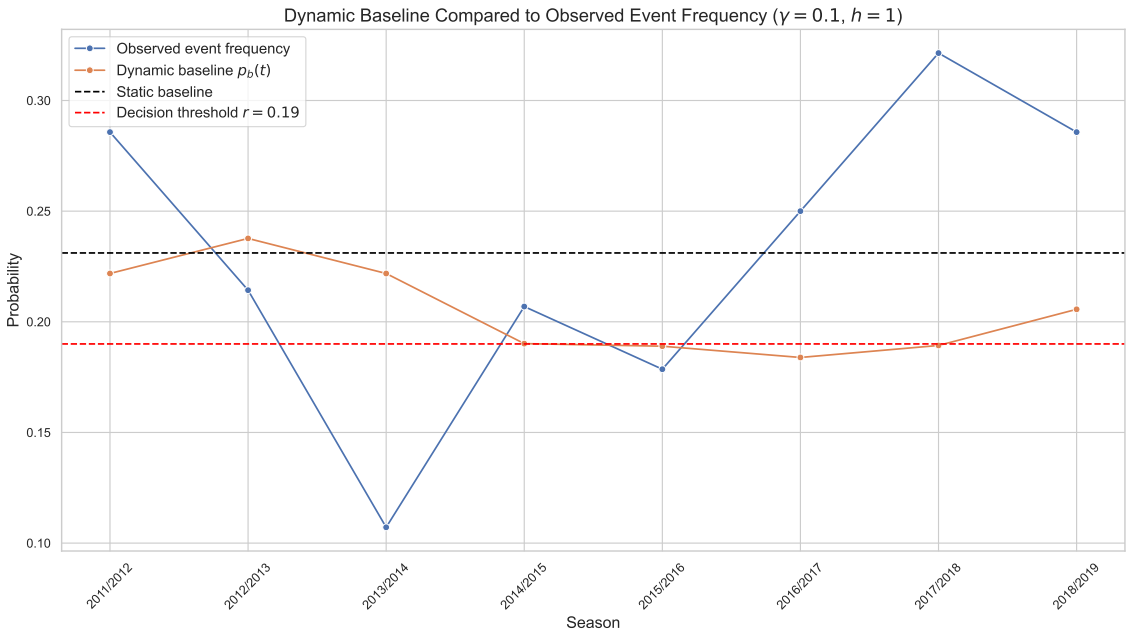

Figure 4.5: Comparison between the true event frequency, the historical baseline probability p_b, the static baseline probability, and the decision threshold $r = 0.19$ for $\gamma = 0.1$ and forecast horizon $h = 1$.

Figure 4.5 shows the observed event frequency for each season (blue line), the dynamic baseline probability p_b (orange line), the decision threshold r (red dashed line), and the static baseline probability computed across all seasons (black dashed line). The Figure is shown for a fixed event definition $\gamma = 0.1$ and forecast horizon $h = 1$.

From Figure 4.5 we observe that the historical baseline probability varies across seasons and lies both above and below the decision threshold r. In contrast, the static baseline probability remains constant across all seasons. Furthermore, the observed event frequency exhibits larger fluctuations between seasons since it's computed each season separately, compared to the historical baseline probability, which varies more gradually over time due to being computed cumulatively across previous seasons.

4.3 Forecast Accuracy

4.3.1 Accuracy performance by horizon

We now move on to evaluate how well the forecasting models perform with respect to the forecast accuracy metrics. All models are evaluated using MAE, LS, and WIS.

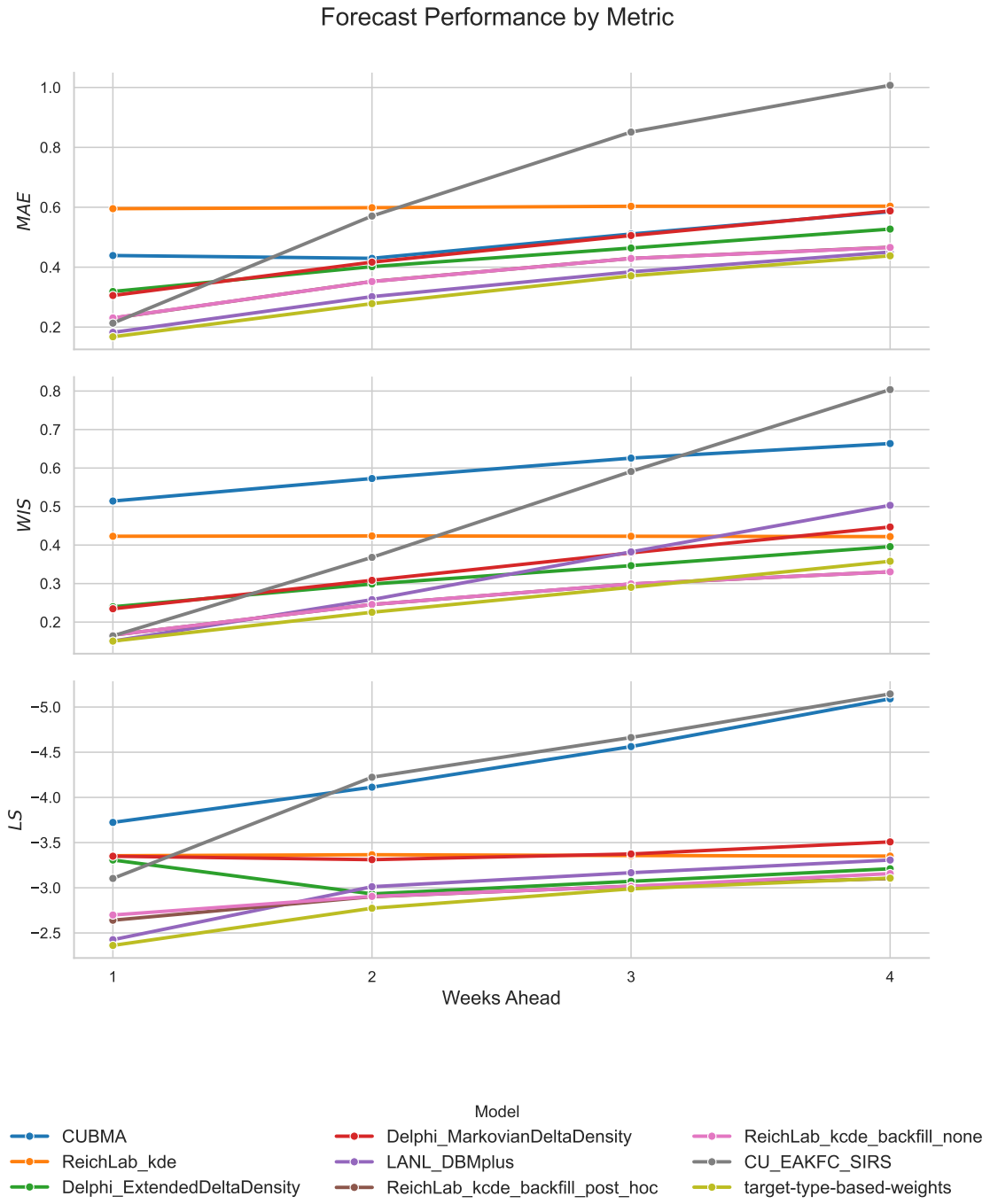

Figure 4.6: Forecast accuracy metrics MAE, LS, and WIS across forecasting models and forecast horizons $h = 1, 2, 3, 4$.

In Figure 4.6 we display the scores of MAE, LS, and WIS for all models across forecast horizons $h = 1, 2, 3, 4$. lower values of MAE and WIS indicate better forecast performance, while higher values of LS indicate better performance. To facilitate comparison between the metrics, the y-axis is flipped for the LS subplot such that lines located lower in the plots correspond to better performing models for all metrics.

Overall, we observe that the ensemble model ”target-type-based-weights” (yellow), alongside the "Reichlab-KCDE" models (pink and brown), generally perform best across the forecast horizons and metrics. Furthermore, the forecast performance generally decreases as the forecast horizon increases. However, the rankings between models differ to some extent depending on the forecast accuracy metric.

4.4 Value of Forecast

4.4.1 Value score curve

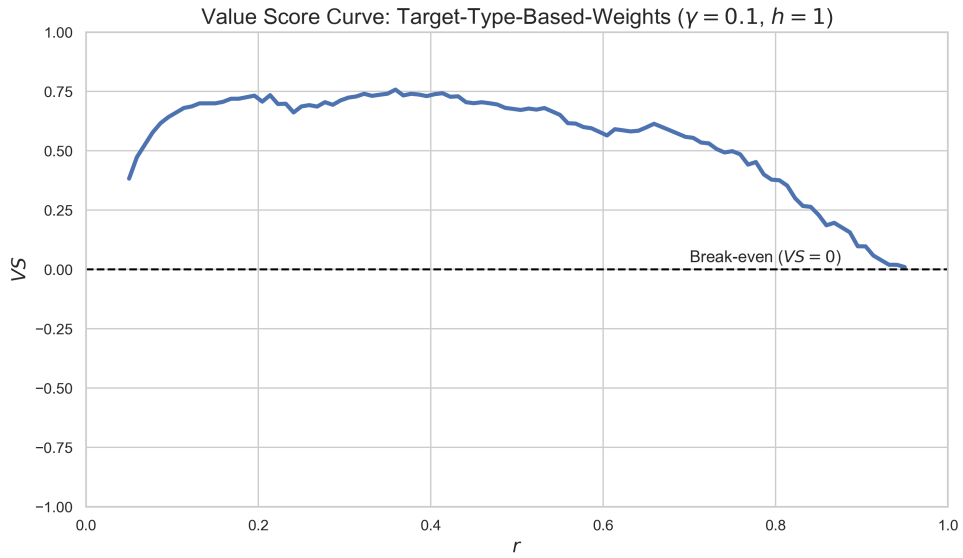

Figure 4.7: Value score as a function of the cost–loss ratio r for the ensemble model “target-type-based-weights” with $\gamma = 0.1$ and forecast horizon $h = 4$. The dashed line corresponds to the break-even value $VS = 0$.

Figure 4.7 shows how the value score varies with respect to the cost-loss ratio for the ensemble model ”target-type-based-weights” with fixed event definition $\gamma = 0.1$ and forecast horizon $h = 4$.

From Figure 4.7, we observe that the value score decreases as the cost–loss ratio increases. However, the value score remains positive across the range of cost-loss ratios and thus stays above the dashed break-even line $VS = 0$, which indicates an improved performance relative to the baseline.

Since the value score curves are parameter-specific, additional summarization is needed to compare model performance across different scenarios.

4.4.2 Integrated decision value

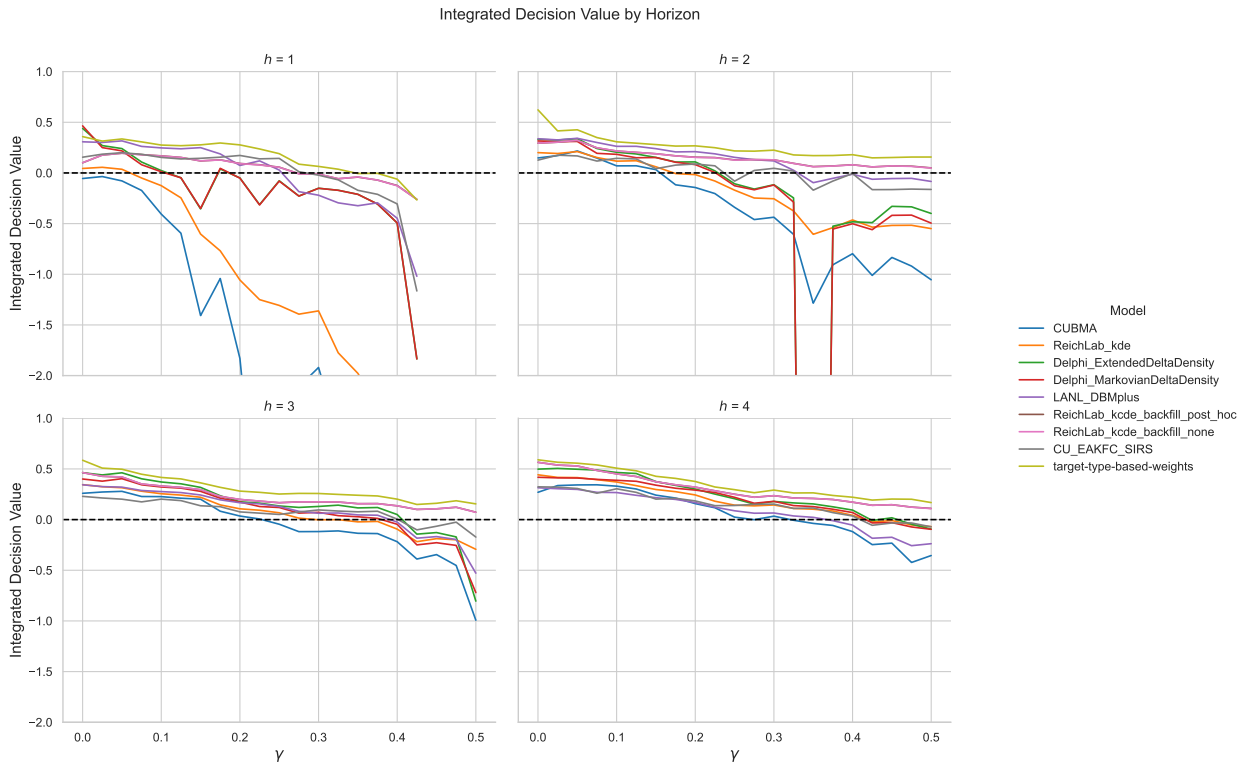

Figure 4.8: Integrated value scores across event definitions γ for all forecasting models and forecast horizons $h = 1, 2, 3, 4$. The dashed line corresponds to the break-even value $IV = 0$.

Figure 4.8 displays the integrated value scores for all forecasting models across different event definitions γ and forecast horizons $h = 1, 2, 3, 4$. Each subplot corresponds to one forecast horizon, while each line represents the integrated value score of one forecasting model obtained by integrating the value score over the cost–loss ratio, as described in Section 3.8. As in the previous section, the dashed black "break even" line, $IV = 0$.

From Figure 4.8, we observe that the integrated value generally decreases as γ increases for all models. Furthermore, the ensemble model ”target-type-based-weights” (yellow), together with the ”Reichlab-KCDE” (pink and brown), generally achieve the highest integrated values across the forecast horizons.

For the forecast horizon $h = 1$, several curves terminate around $\gamma = 0.4$. This occurs because the denominator in the value score becomes zero for certain parameter combinations, making the integrated value undefined.

4.4.3 Decision value surface

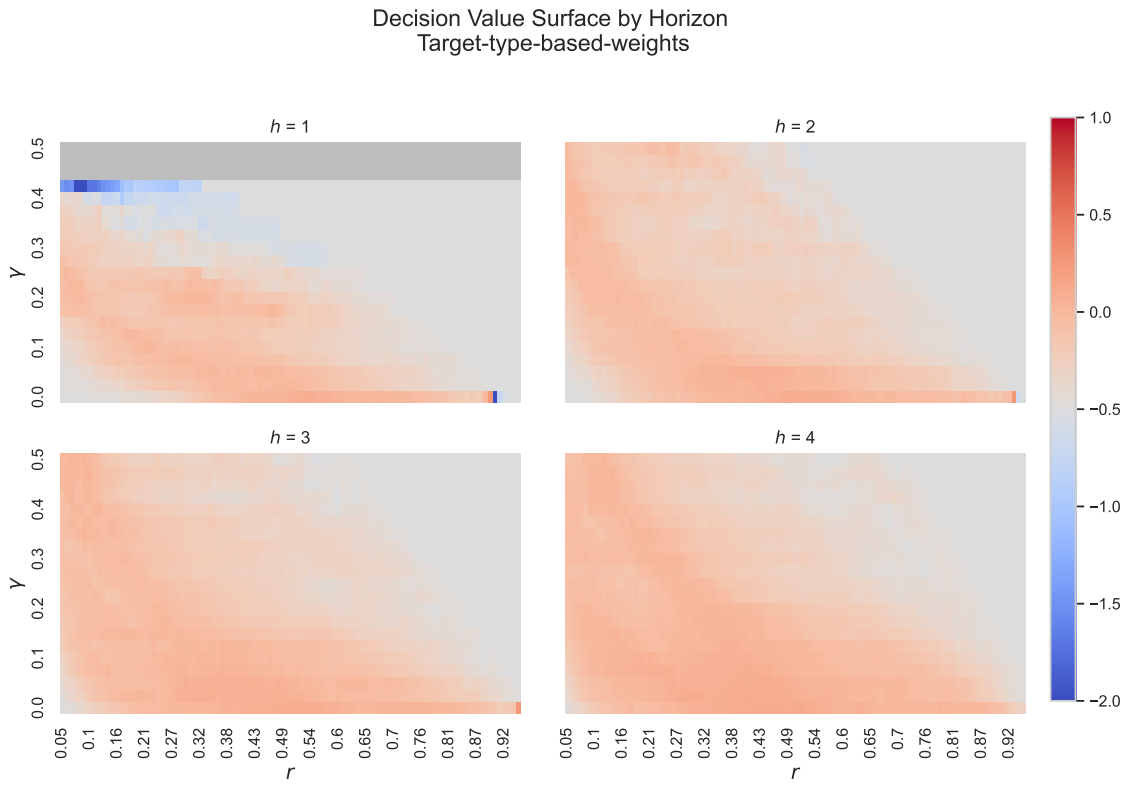

Figure 4.9: Value score surfaces for the ensemble model "target-type-based-weights" across forecast horizons $h = 1, 2, 3, 4$. The heatmaps display the value score as a function of the cost–loss ratio r and the event definition parameter γ.

Figures 4.9 and 4.10 display the value score surfaces for the ensemble model 'target-type-based-weights' and "CUBMA" across forecast horizons h. Each subplot corresponds to one forecast horizon, where the heatmaps display the value score as a function of the cost–loss ratio r (x-axis) and the event definition parameter γ (y-axis).

As mentioned in the previous section, the forecast horizon $h = 1$ contains dark grey regions corresponding to undefined values caused by division by zero in the value score computation.

Overall, the ensemble model "target-type-based-weights" exhibits mostly positive or near-zero value scores across most parameter combinations. In contrast, the "CUBMA" model exhibits mostly negative value scores across most parameter combinations.

Furthermore, the Value Scores generally decrease as both the cost–loss ratio r and the event definition parameter γ increase, producing an approximately triangular pattern across the heatmaps.

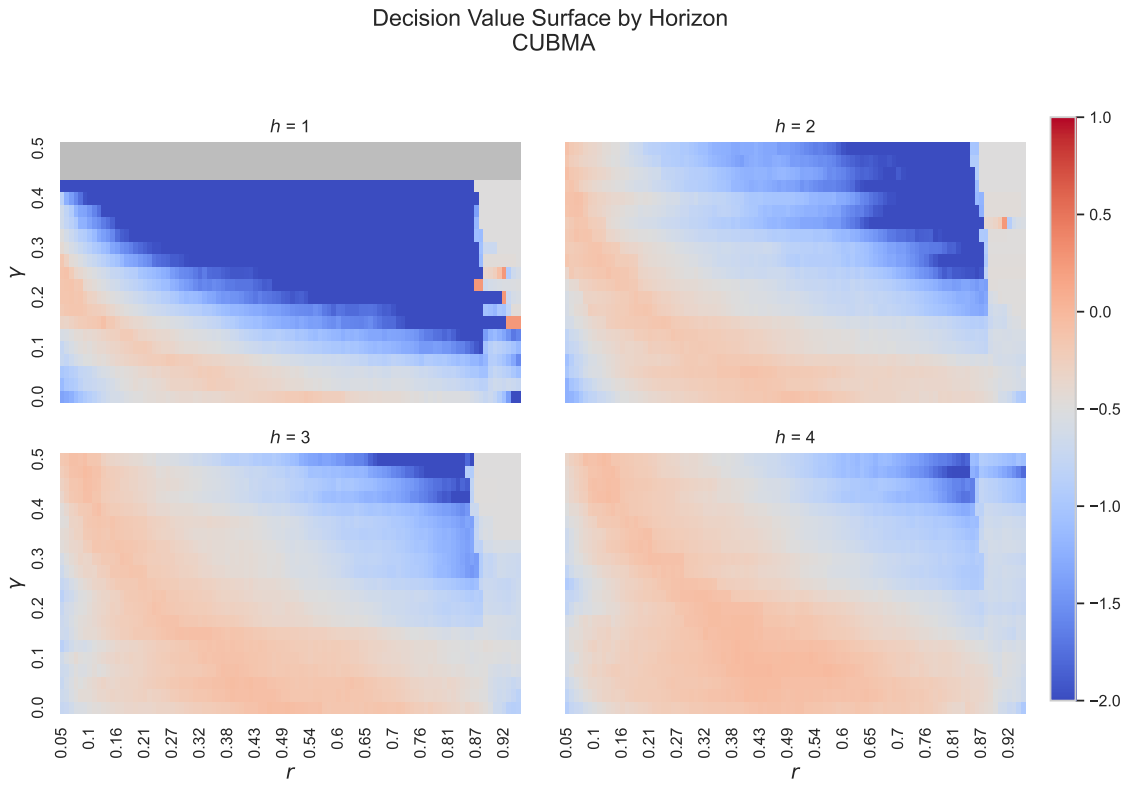

Figure 4.10: Value score surfaces for the "CUBMA" model across forecast horizons $h = 1, 2, 3, 4$. The heatmaps display the value score as a function of the cost–loss ratio r and the event definition parameter γ.

4.5 Rank Correlation

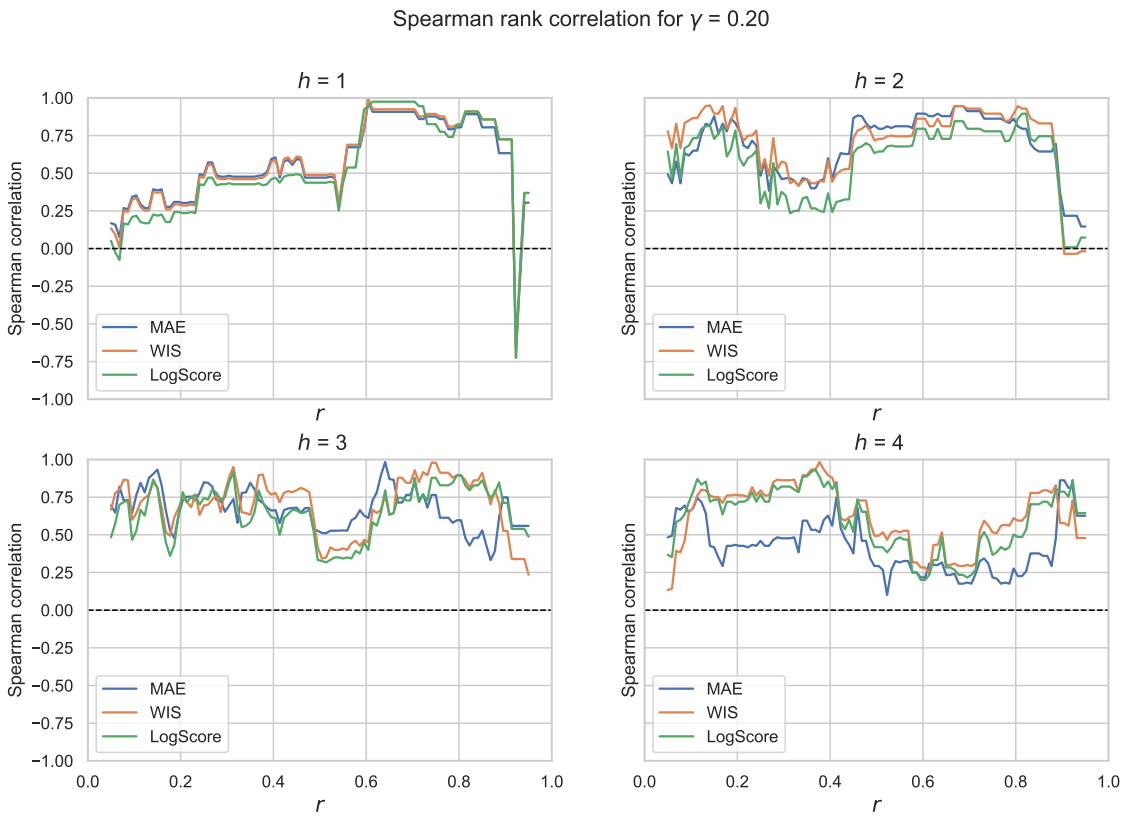

Figure 4.11: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.20$ across forecast horizons and cost–loss ratios.

Lastly, we compare the rankings induced by the forecast accuracy metrics with the rankings induced by the value of forecast framework. Figures 4.11 and 4.12 display the Spearman rank correlation between the forecast accuracy rankings and Value Score rankings for event definitions $\gamma = 0.2$ and 0.45. Each Figure consists of subplots corresponding to the forecast horizons $h = 1, 2, 3, 4$, where each subplot contains three lines representing the forecast accuracy metrics MAE, LS, and WIS being compared to VS across the cost-loss ratios r.

Overall, the rank correlations follow relatively similar patterns across the forecast accuracy metric, although they differ somewhat in absolute magnitude across the cost–loss ratios. Furthermore, the rank correlations are mostly positive in Figures 4.11 and 4.12, while negative rank correlations occur relatively infrequently.

The rank correlation for the remaining γ values is provided in Appendix A for event definitions ranging from $\gamma = 0$ to $\gamma = 0.5$. For forecast horizon $h = 1$ and event definitions $\gamma \geq 0.45$, no rank correlation curves are displayed due to undefined Value Scores.

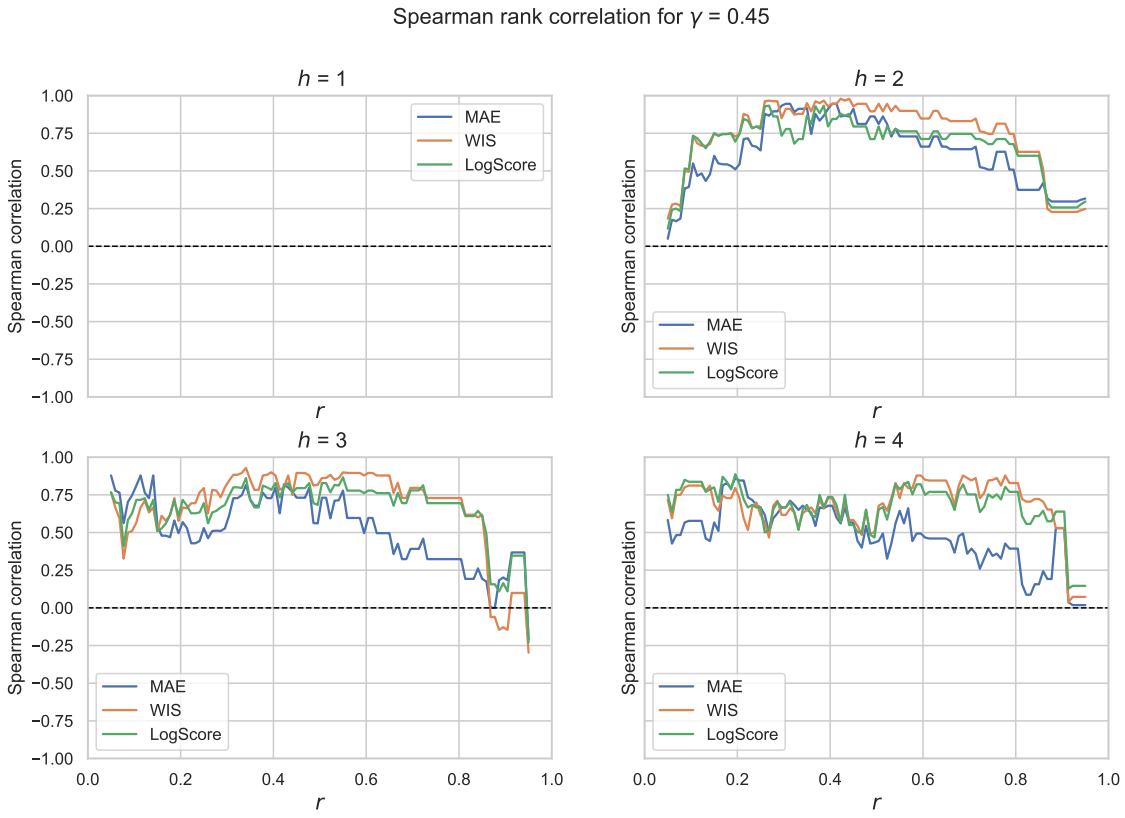

Figure 4.12: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.45$ across forecast horizons and cost–loss ratios.

5
Discussion

In this chapter, the results presented are interpreted and discussed in relation to the research questions addressed in this thesis. First, the relationship between forecast accuracy and forecast value is examined. Then the discussion considers the influence of event definitions, forecast horizons, and decision settings, before concluding with the limitations of the study and suggesting future research areas.

5.1 Forecast Accuracy vs Decision Value

Overall, as shown in Section 4.5, the rankings induced by forecast accuracy metrics and the value of forecast framework are generally positively correlated across most cost–loss ratios and event definitions, see Figures 4.11, 4.12 and the remaining Figures in Appendix A. Although the correlation magnitude varies, the rankings frequently exhibit moderate to strong positive agreement. This suggests that the different evaluation frameworks often agree on which forecasting models perform relatively well or poorly. Consequently, models with strong forecast accuracy frequently also exhibit high decision value.

This behavior is also reflected in the model-specific results. From Figure 4.8, the ensemble model ”target-type-based-weights” together with the ”Reichlab-KCDE” models achieve the highest integrated values across forecast horizons. These observations are consistent with the forecast accuracy results shown in Figure 4.6, where these models also rank among the strongest performers. Similarly, the "CUBMA" model performs relatively poorly under both evaluation approaches.

Furthermore, no clear systematic pattern can be observed with respect to the event definition γ, the cost–loss ratio r, or the forecast horizon h. Although the rank correlations vary across parameter combinations, the overall behavior remains relatively similar. Changing the event definition, the cost–loss ratio, or varying the forecast horizon does not appear to consistently strengthen or weaken the agreement between the ranking systems. This suggests that the relationship between forecast accuracy and decision value remains relatively stable across the considered decision settings.

However, the agreement between the ranking systems is not perfect. For example, the model ”CU_EAKFC_SIRS” appears to achieve relatively stronger performance under the value of forecast framework compared to its forecast accuracy ranking. This suggests that models with similar statistical accuracy can differ in their useful-

ness from a decision-making perspective. Since forecast accuracy evaluates closeness to observed outcomes while the value of forecast framework evaluates decision performance under a cost–loss setting, different aspects of forecast quality are emphasized. Consequently, strong forecast accuracy does not necessarily guarantee optimal decision value.

Overall, these observations indicate that the relationship between accuracy and decision value is not entirely straightforward. Although the ranking systems frequently agree, deviations between them suggest that forecast accuracy alone may not fully characterize model usefulness under a decision-making setting.

5.2 Effect of Event Severity and Horizon

When increasing the event definition parameter γ, Figure 4.3 shows that the event frequency decreases. In other words, larger event definitions correspond to increasingly rare events. This effect appears stronger for smaller forecast horizons.

This behavior also explains why Figures 4.9 and 4.10 contain undefined values for forecast horizon $h = 1$ and larger values of γ. As outlined in Algorithm 1, the Value Score computation relies on expected expenses computed from observed forecast instances. When the event becomes sufficiently rare, there may be too few or no observed instances supporting the calculations. In practice, this creates situations where all decision settings behave similarly because the event rarely occurs. Consequently, the Value Score becomes undefined.

This observation is supported by Figure 4.2, where shorter forecast horizons exhibit tighter growth rate distributions. In contrast, larger forecast horizons show wider distributions, which allow for more substantial growth events. As a result, larger values of γ remain observable for longer horizons.

Therefore, for each forecast horizon, there exists a sufficiently large event definition γ for which the decision settings eventually become indistinguishable. As γ increases, the event becomes increasingly rare, and all models converge toward similar behavior since little or no event information remains available. The difference is that larger horizons are capable of supporting larger values of γ due to their broader growth rate distributions.

Furthermore, Figure 4.6 shows that forecast accuracy generally deteriorates as the forecast horizon increases. This behavior is expected since uncertainty increases further into the future. Interestingly, Figure 4.8 suggests that larger horizons produce larger integrated value scores. Although forecasts become statistically less accurate, the broader range of possible future outcomes, as we've seen in Figure 4.2, may create additional opportunities in which forecast information is useful for decision-making. Another possibility is that the baseline model becomes relatively worse for larger horizons.

Moreover, wider growth rate distributions for larger forecast horizons are expected since predictions further into the future naturally involve greater uncertainty. Larger horizons allow more time for changes in ILI prevalence and therefore enable a wider range of possible growth scenarios. This behavior is consistent with general forecasting intuition, where uncertainty typically accumulates further into the future.

Figures 4.9 and 4.10 further illustrate how the value score behaves jointly across the parameters r, h, and γ. Increasing the cost-loss ratio generally appears to be associated with lower Value Scores. One possible explanation is that as r approaches one, the cost of preventive becomes increasingly similar to the loss incurred by inaction. Consequently, the distinction between decisions becomes smaller, and the benefit of forecast information may decrease.

Similarly, increasing γ creates increasingly restrictive event definitions. Since larger event definitions correspond to rarer events, less historical information becomes available for evaluating model performance. Together, these effects create a pattern observed in the decision value surfaces.

The lower-left regions of the decision surfaces, corresponding to small values of γ and r, also display lower value scores. In such settings, actions become very inexpensive, and events become common, leading to situations in which action is frequently taken regardless of the forecast, since only a small forecast probability p_f is required to exceed the cost–loss ratio threshold. Conversely, when γ becomes very large, events become so rare that limited information remains available for distinguishing models. Therefore, the forecasting models appear most useful within intermediate regions where event definitions and decision settings are sufficiently balanced.

5.3 Interpretation of the Cost–Loss Framework

This thesis extends traditional forecast evaluation by incorporating a framework that accounts for different decision settings. Figure 4.7 illustrates how the Value Score varies across cost-loss ratios for a fixed model, forecast horizon, and event definition. Such curves provide useful information for a specific decision context but remain limited in the amount of information they convey.

To extend this perspective, Figures 4.9 and 4.10 introduce an additional dimension by displaying how decision value varies jointly across cost-loss ratios and event definitions. These visualizations provide a broader overview of model behavior under different decision scenarios.

However, while decision surfaces provide a detailed view of a model's performance across event definitions and cost–loss ratios, they can be difficult to compare across many forecasting models. Since each model gives rise to a separate heatmap, the number of visualizations quickly grows large even with a small number of forecasting models. Furthermore, the two-dimensional nature of heatmaps can make it

challenging to obtain an overall understanding of a model's performance. Therefore, integrated value scores were introduced to summarize model performance across the considered decision range of cost-loss ratios. By integrating over this range, we obtain a lower-dimensional representation that displays the overall usefulness of a model.

However, this approach relies on the assumption that the chosen range of cost–loss ratios represents meaningful decision settings. Under this assumption, the integrated value score can be interpreted as a measure of how useful a forecasting model is across a broader set of decision scenarios rather than within a single specific context. The choice of integration range is therefore important, since different ranges correspond to emphasizing different decision settings. In the special case where the Value Score is aggregated across the full range of cost-loss ratios, Murphy [5] demonstrated that this is related to the Brier Score, which is a commonly used measure for probabilistic forecast accuracy.

5.4 Baseline Construction and Historical Information

When constructing the baseline, this thesis focused on a dynamic approach where only information from previous seasons is used when computing the baseline probability and thus the value score. As shown in Figure 4.5, the static baseline probability computed across all seasons is relatively close to the dynamic baseline.

However, Figure 4.5 also illustrates an important advantage of the dynamic baseline. Since the baseline probability evolves, the relationship between the decision threshold and the historical baseline may change throughout the seasons. This creates a more realistic setting where beliefs are updated continuously, and it avoids information leakage from future observations.

The dynamic baseline nevertheless has potential drawbacks. Unexpected seasons or substantial shifts in the disease behavior could temporarily make the baseline less representative. However, Figures 4.1 and 4.4 suggest that although seasonal variation exists, most seasons exhibit relatively similar overall behavior. Furthermore, the expanding mean construction reduces the influence of individual seasons.

5.5 Limitations

Although this thesis provides insight into how the forecast accuracy relates to the value of forecast framework, several limitations should be acknowledged.

Firstly, rank correlation analysis is based on only nine forecasting models. Given the relatively small sample of models, the resulting Spearman rank correlations may be sensitive to rank changes and should be interpreted with some caution. Fur-

thermore, the analysis focuses on the agreement between framework rankings rather than the stability of individual model ranks across decision settings. Consequently, questions such as whether certain models consistently perform well across decision settings aren't investigated in detail except for the integrated value score.

Secondly, the analysis relies on a binary event formulation, where an event either occurs or doesn't occur. This represents a simplification of real-world decision settings, where outcomes and actions may involve several discrete levels of intervention or even a continuous interval of interventions. An article by Murphy discusses generalized decision settings extending beyond binary events and actions, [6]. Extending the proposed framework to include such settings would thus be a natural continuation for future research.

Furthermore, the event definition used in this thesis (2.5) represents only one possible way of constructing events. One can construct alternative event definitions that emphasize different aspects of disease prevalence, which potentially can alter the relationship between forecast accuracy and decision value. An alternative event formulation is explored in [7]. [7] considers an event definition based on whether the seasonal influenza peak intensity exceeds a predefined severity threshold. In that setting, the value of forecast framework evaluates forecasts according to their ability to predict severe influenza seasons, rather than the ability to predict short-term relative increases in ILI. Consequently, different event constructions may therefore emphasize different aspects of forecast performance and lead to different assessments of forecast value.

Another limitation is that the analysis focused exclusively on events corresponding to increases in influenza activity. Analogous event definitions corresponding to decreases in disease prevalence could be constructed, however their interpretation within the cost-loss framework is less straightforward. When influenza activity increases, it naturally corresponds to situations where preventive actions such as increasing staffing levels for example may reduce future expenses. On the other hand, for decreasing influenza activity, the corresponding actions and associated costs and losses are less clearly defined.

Finally, this thesis focuses on one forecasting dataset based on ILI data. Although Influenza forecasting provides a suitable setting for studying probabilistic forecasting in epidemiology, more recently, forecast prediction in epidemiology may involve additional targets than ILI, and disease prediction can also be done on other diseases, such as COVID-19.

5.6 Future Research

One possible extension of this work would be to investigate alternative approaches for aggregating value scores across cost-loss ratios. In this thesis, all decision settings are treated equally when computing the integrated value. However, some decision contexts may be more relevant than others, depending on the preference of a decision

maker. Therefore, introducing weighted cost–loss ratios could provide a more realistic representation of practical decision-making settings regarding the integrated value score.

Future work may also consider more complex decision settings involving multiple actions and event categories, rather than binary decisions. Furthermore, applying the framework to additional diseases, newer forecasting systems, or alternative event definitions may provide further insight into the relationship between forecast accuracy and decision value.

6
Conclusion

Influenza forecasts have historically often been evaluated through statistical accuracy measures. However, as shown throughout this study, strong forecast accuracy does not necessarily imply decision usefulness. The objective of this thesis has therefore been to introduce and investigate a cost-loss framework for evaluating the value of forecast in comparison with traditional forecast accuracy metrics. Furthermore, we investigated both frameworks individually and the relationship between their induced ranking systems using Spearmam rank correlation. This was studied using probabilistic forecasts from FluSight data, where uncertainty in future predictions is represented explicitly through predictive distributions.

The cost–loss framework was implemented by defining events based on the growth between current and future observations across forecast horizons. These event definitions generated forecast probabilities, which were then compared against cost–loss ratios acting as decision thresholds. This created a decision process where forecasts contributed to expected expenses based on forecast, historical baseline, and perfect foresight scenarios. These quantities were subsequently combined into Value Scores depending on event definition, forecast, horizon, and cost loss ratio. To facilitate comparison between forecasting models, integrated value measures and rank correlations were further introduced.

The findings indicate that the forecast accuracy rankings and value of forecast rankings are generally positively correlated across event definitions, cost-loss ratios, and forecast horizons. However, the relationship is not perfectly aligned. Although the ensemble model, ”target-type-based-weights”, and ”Reichlab-KCDE” performed strongly under both frameworks, certain differences between the rankings systems were still observed.

As a final remark, this thesis suggests that the forecast accuracy and value of forecast are largely aligned within the considered setting and its limitations. However, the value of forecast framework still provides additional insight by allowing forecast usefulness to be interpreted within specific decision contexts. This introduces information that traditional forecast accuracy measures alone are unable to capture.

Bibliography

[1] C. Mills, N. Irons, J. Tsui, S. Sparrow, L. Carvalho, A. Kucharski, O. Ratmann, B. Lambert, C. Donnelly, and M. Kraemer, "From metric to action: An evaluation framework to translate infectious disease forecasts into policy decisions," Jul. 2025, medRxiv preprint.

[2] N. G. Reich, C. J. McGowan, T. K. Yamana, A. Tushar, E. L. Ray, D. Osthus, S. Kandula, L. C. Brooks, W. Crawford-Crudell, G. C. Gibson, E. Moore, R. Silva, M. Biggerstaff, M. A. Johansson, R. Rosenfeld, and J. Shaman, "Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the u.s." *PLoS Computational Biology*, vol. 15, no. 11, p. e1007486, 2019. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1007486

[3] 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, pp. 1–15, Feb. 2021. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1008618

[4] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007. [Online]. Available: https://doi.org/10.1198/016214506000001437

[5] A. H. Murphy, "The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation," *Monthly Weather Review*, vol. 105, no. 7, pp. 803–816, 1977.

[6] ——, "Decision making and the value of forecasts in a generalized model of the cost-loss ratio situation," *Monthly Weather Review*, vol. 113, no. 3, pp. 362–369, 1985.

[7] P. Gerlee, T. Lundh, A. S. Jöud, and H. Thorén, "Evaluating infectious disease forecasts in a cost-loss situation," 2026. [Online]. Available: https://arxiv.org/abs/2601.05921

[8] World Health Organization, "Global epidemiological surveillance standards for influenza," 2013, accessed: 2026-04-30. [Online]. Available: https://apps.who.int/iris/handle/10665/311268

[9] N. G. Reich, L. C. Brooks, S. J. Fox, S. Kandula, C. J. McGowan, E. Moore, D. Osthus, E. L. Ray, A. Tushar, T. K. Yamana, M. Biggerstaff,

M. A. Johansson, R. Rosenfeld, and J. Shaman, "A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the united states," *Proceedings of the National Academy of Sciences*, vol. 116, no. 8, pp. 3146–3154, 2019. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6386665/

[10] A. Gerding, N. G. Reich, B. Rogers, and E. L. Ray, "Evaluating infectious disease forecasts with allocation scoring rules," *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 188, no. 4, pp. 1299–1325, 10 2025. [Online]. Available: https://doi.org/10.1093/jrsssa/qnae136

[11] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

[12] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[13] FluSightNetwork, "cdc-flusight-ensemble," https://github.com/FluSightNetwork/cdc-flusight-ensemble, 2026, accessed April 27, 2026.

[14] cdcepi, "Flusight-forecast-hub," https://github.com/cdcepi/FluSight-forecast-hub, 2026, accessed April 27, 2026.

[15] N. G. Reich, C. J. McGowan, T. K. Yamana, A. Tushar, E. L. Ray, D. Osthus, S. Kandula, L. C. Brooks, W. Crawford-Crudell, G. C. Gibson, E. Moore, R. Silva, M. Biggerstaff, M. A. Johansson, R. Rosenfeld, and J. Shaman, "Supplement (s1 text) for "accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the u.s."," Nov. 2019, supplementary material dated November 5, 2019.

A

Appendix 1

This appendix contains the remaining Spearman rank correlation figures between the forecast accuracy rankings and the value score rankings for all event definitions $\gamma \in [0, 0.5]$. Each figure consists of subplots corresponding to forecast horizons $h = 1, 2, 3, 4$, where the lines represent the forecast accuracy metrics MAE, LS, and WIS across cost–loss ratios r.

For forecast horizon $h = 1$ and event definitions $\gamma \geq 0.45$, no rank correlation curves are displayed due to undefined value scores.

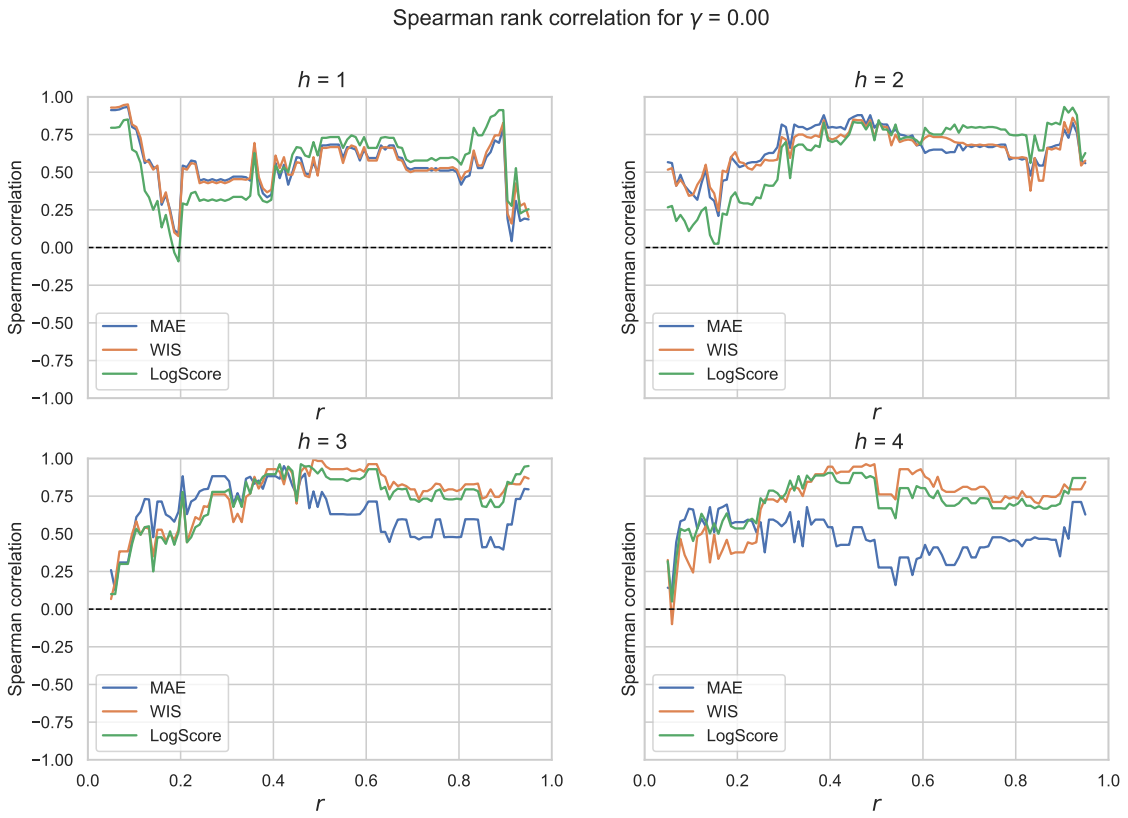

Figure A.1: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.00$.

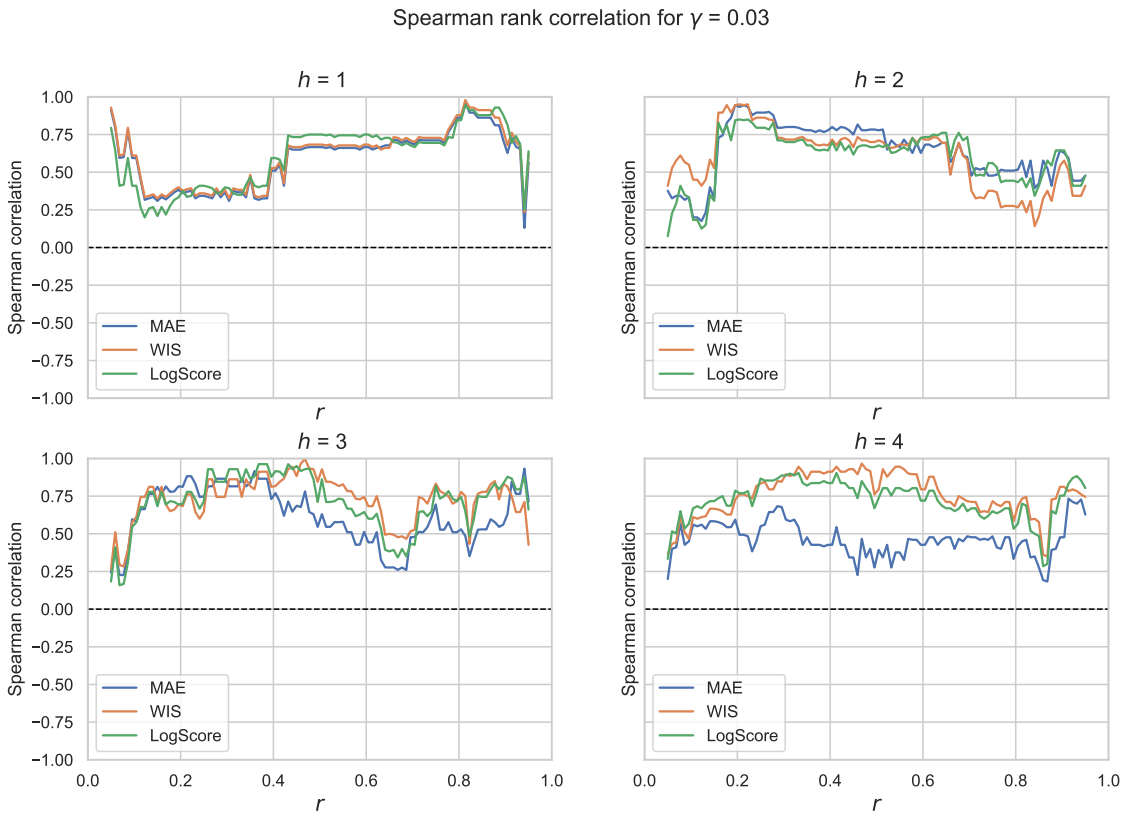

Figure A.2: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.03$.

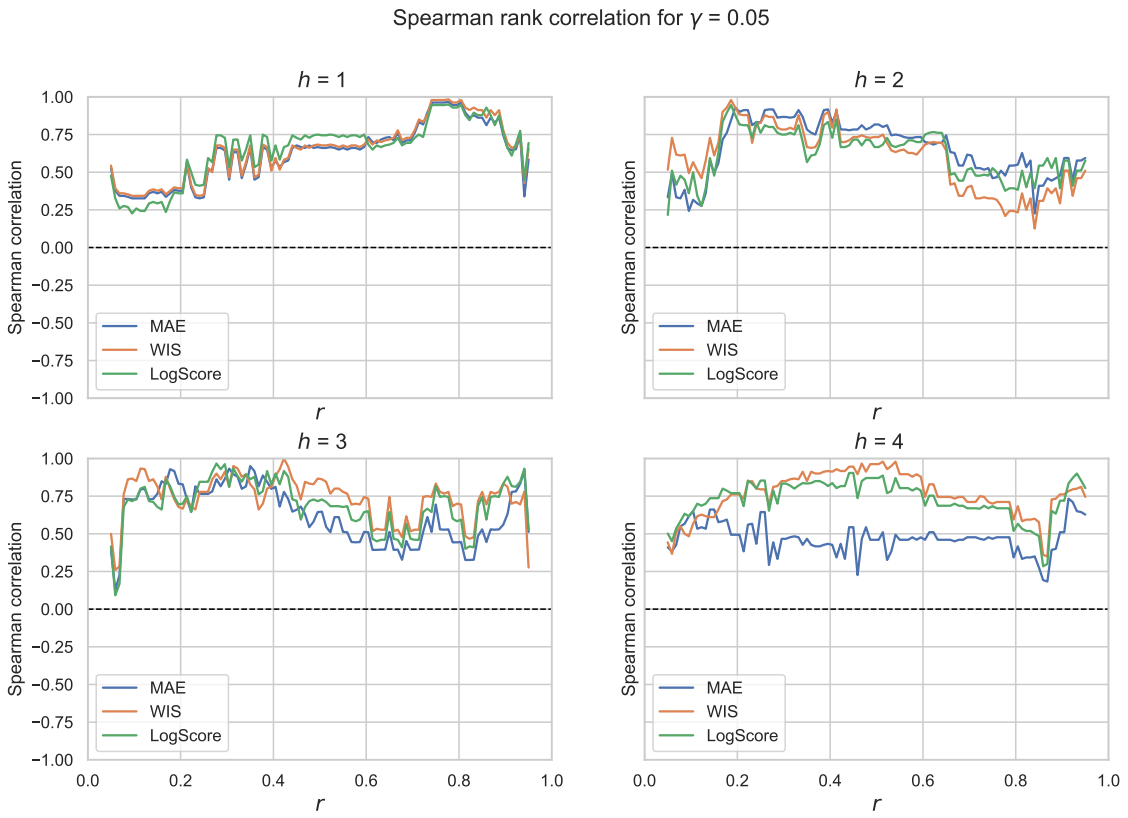

Figure A.3: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.05$.

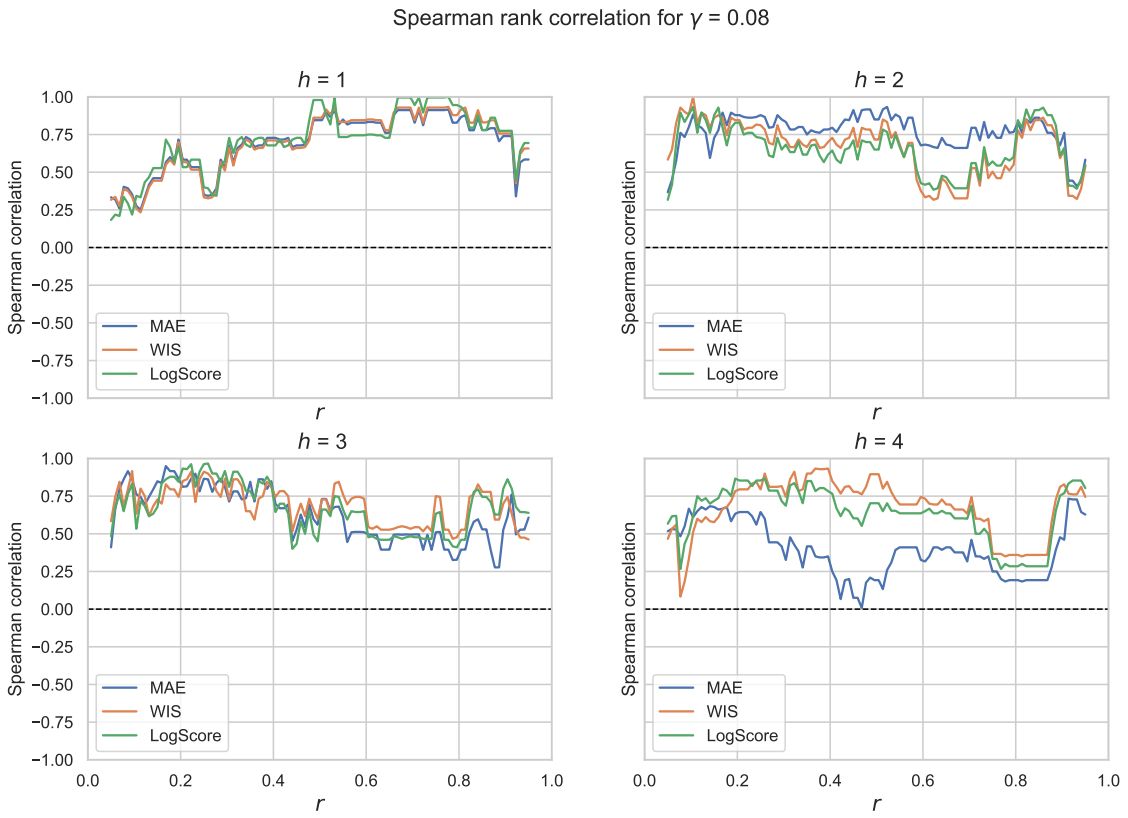

Figure A.4: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.08$.

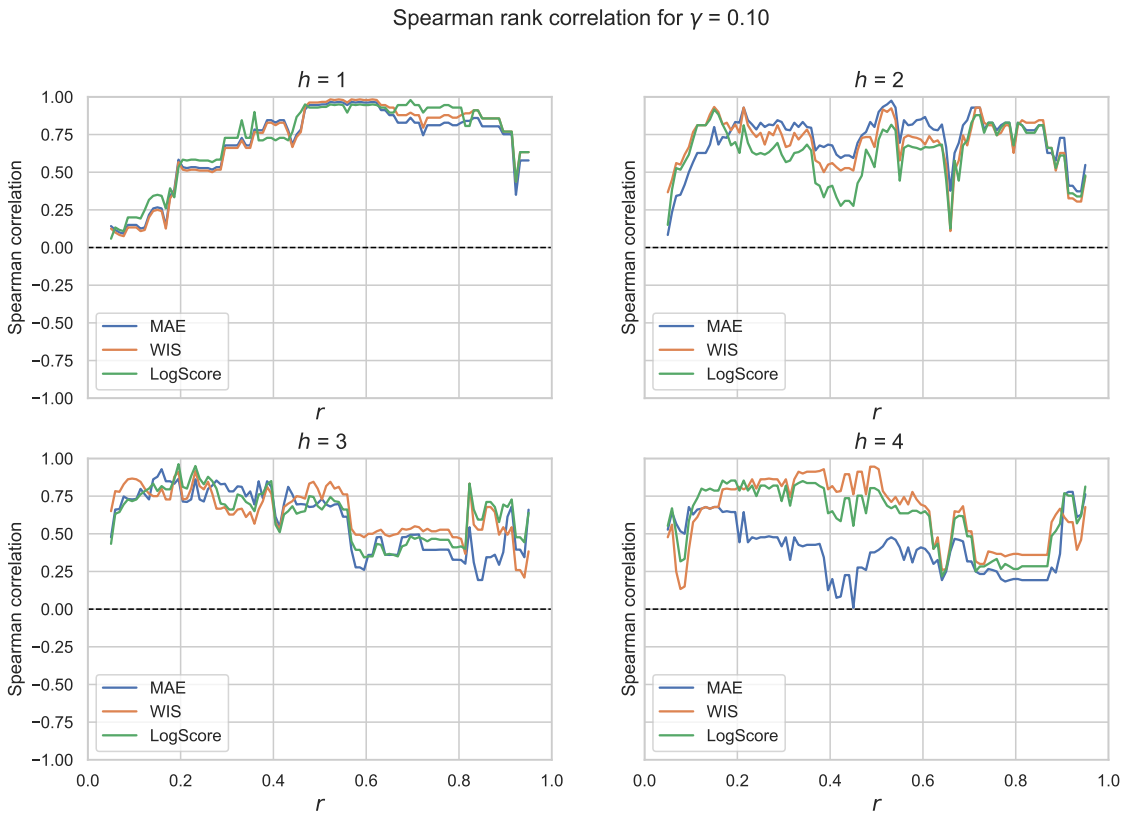

Figure A.5: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.10$.

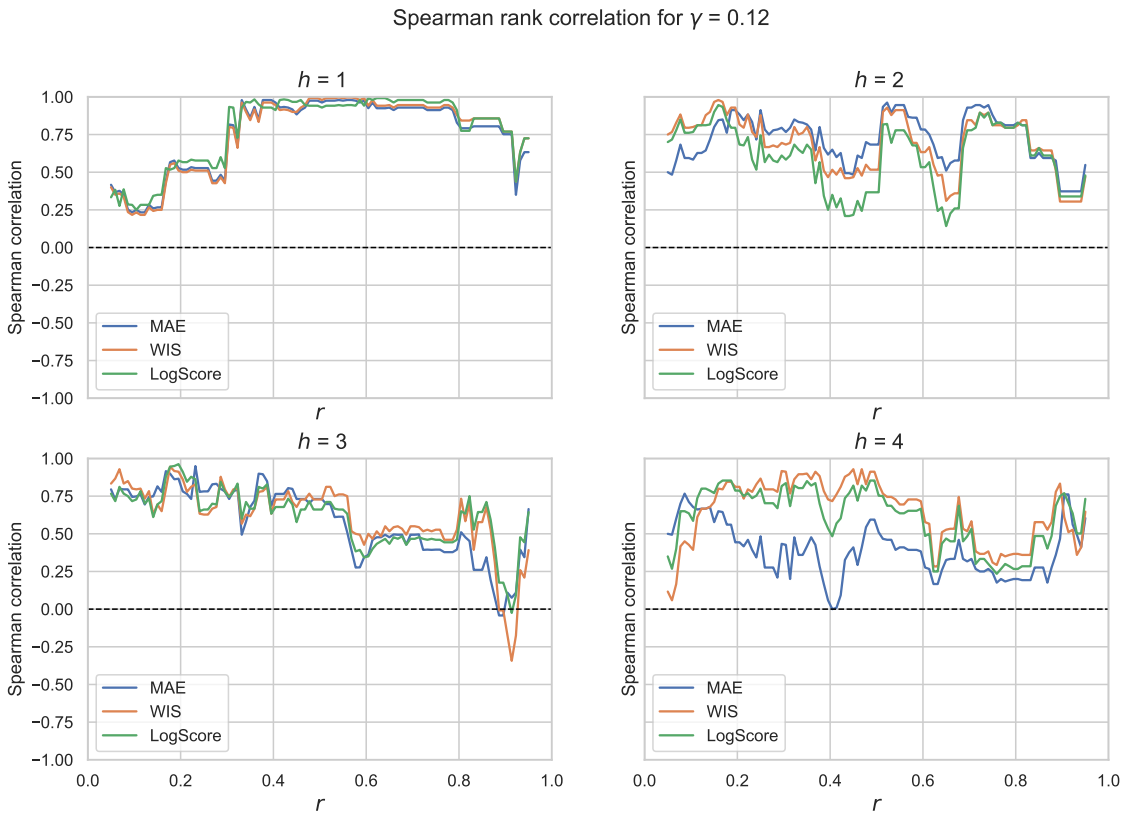

Figure A.6: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.12$.

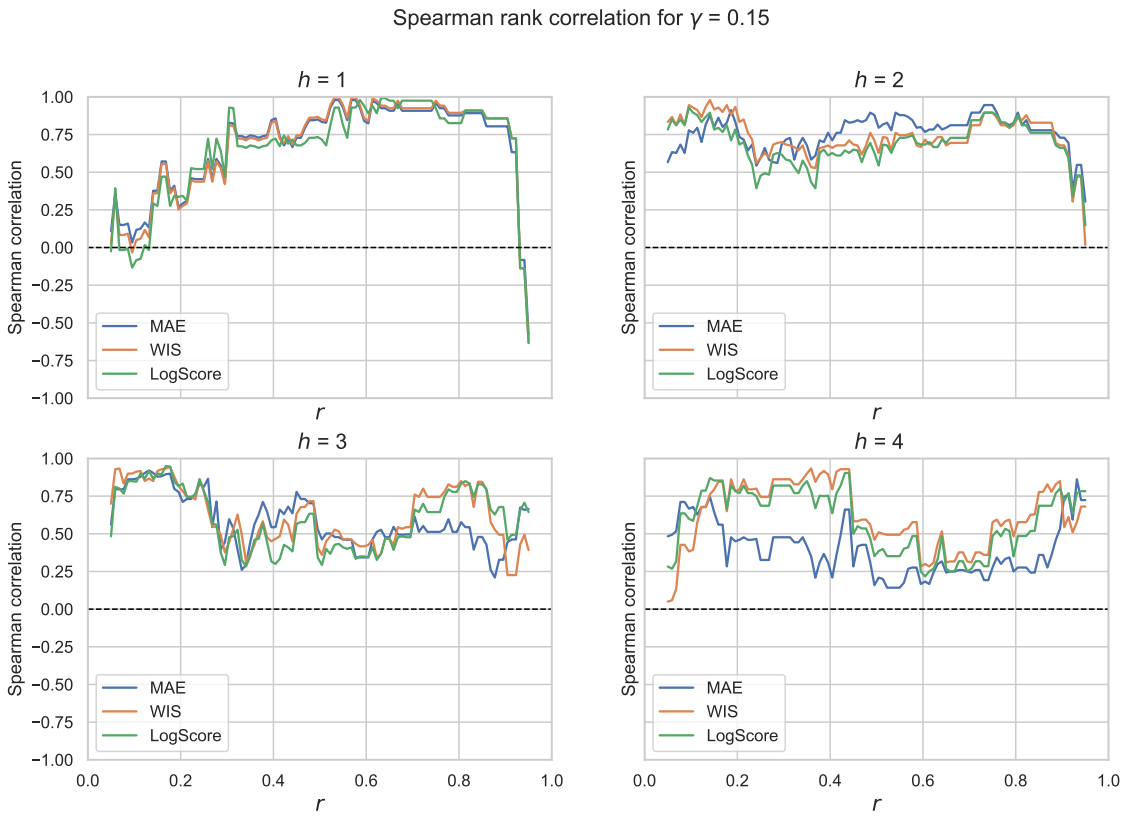

Figure A.7: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.15$.

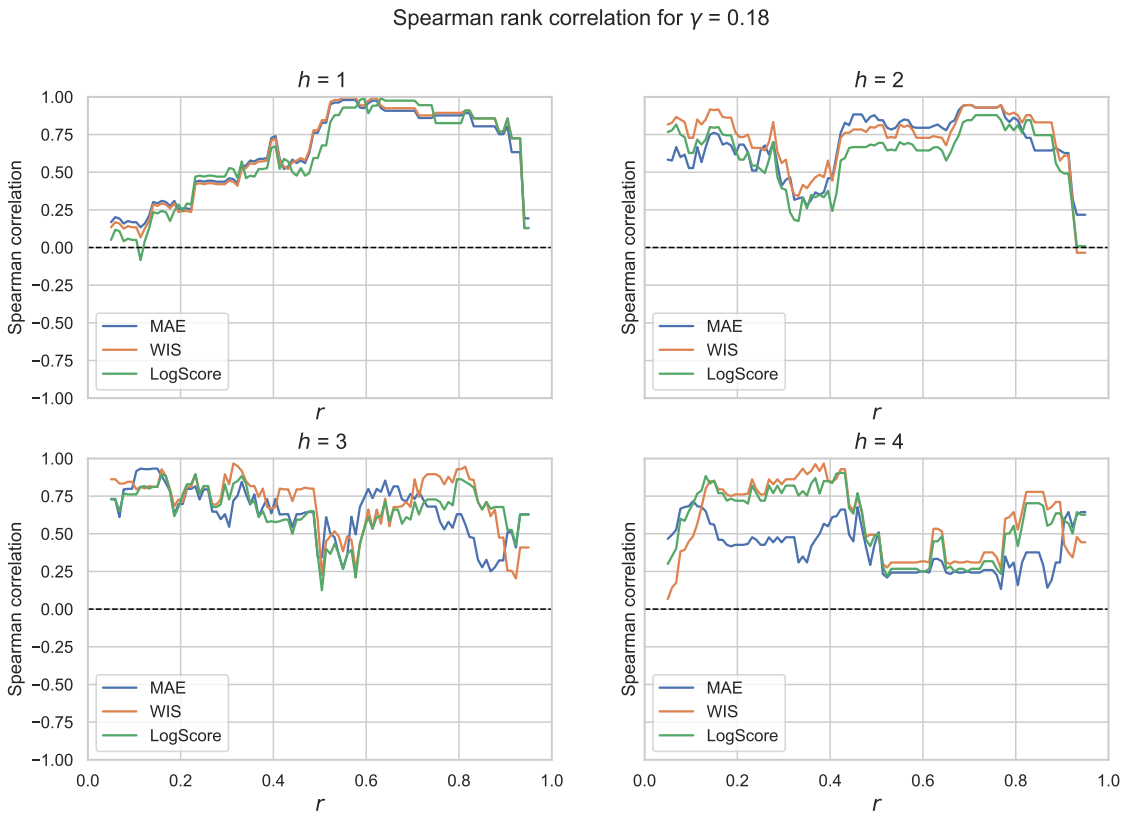

Figure A.8: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.18$.

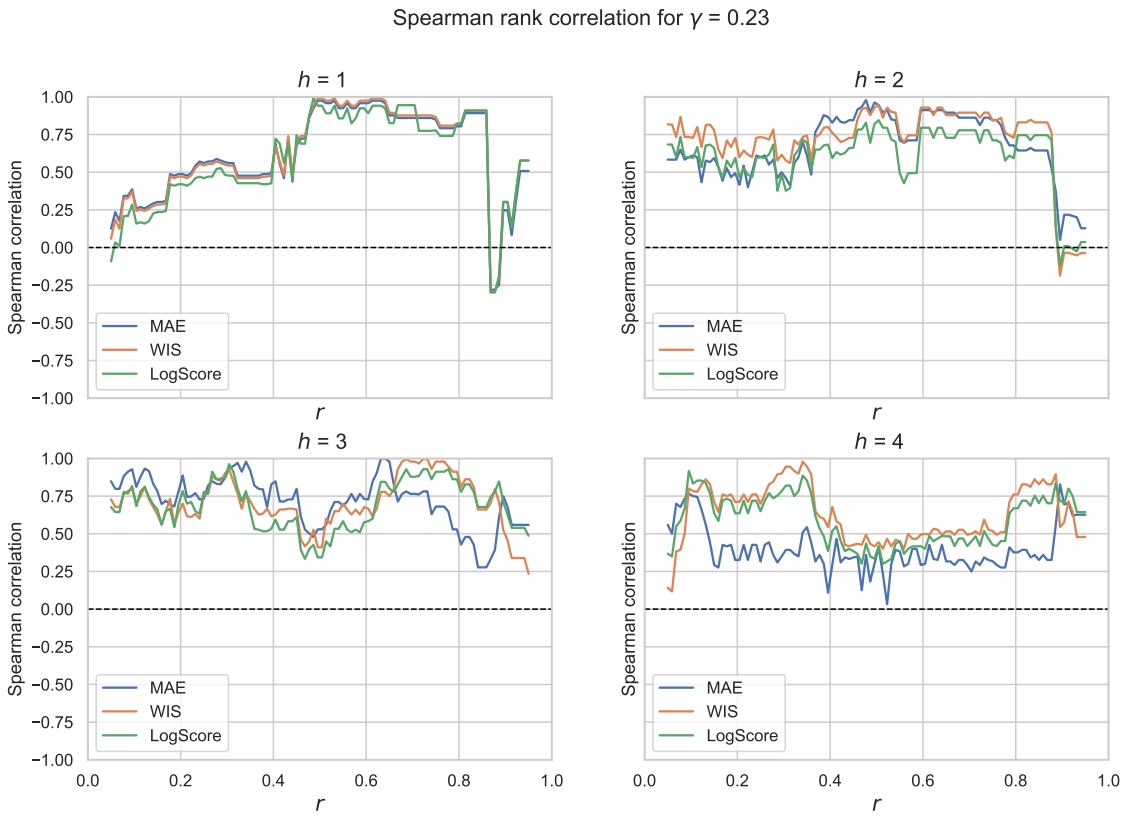

Figure A.9: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.23$.

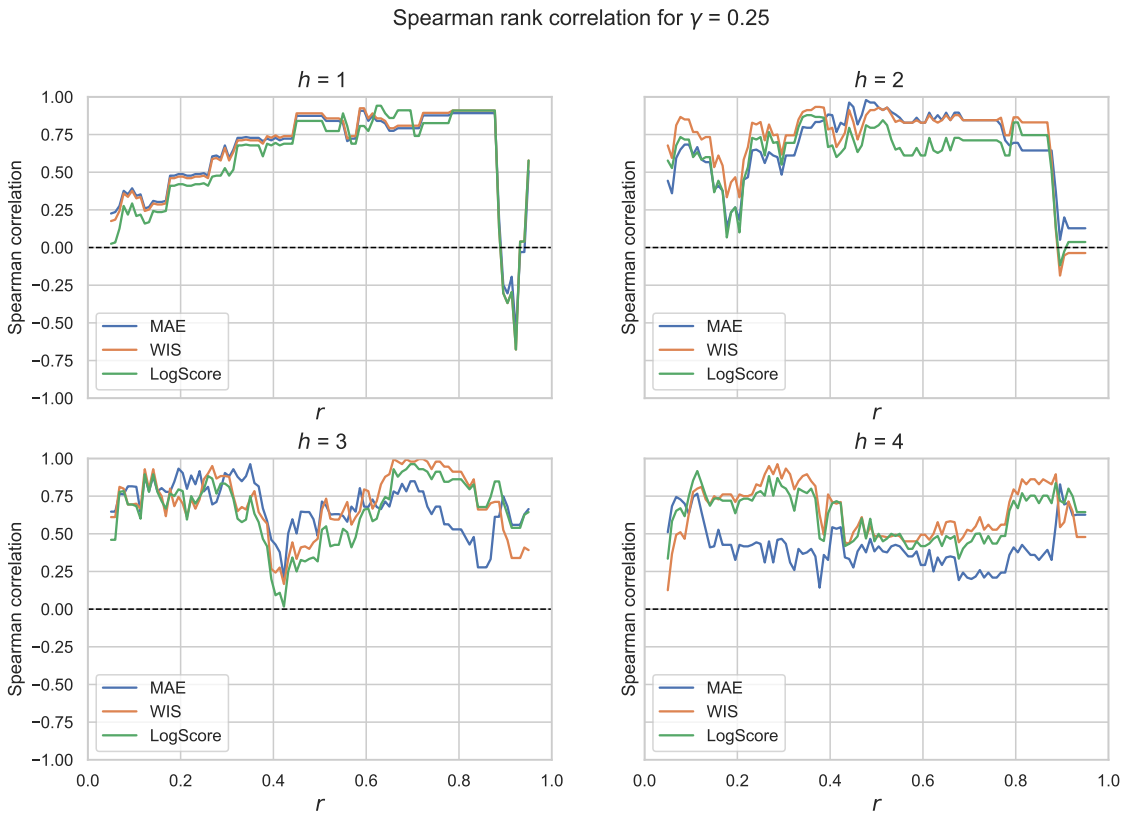

Figure A.10: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.25$.

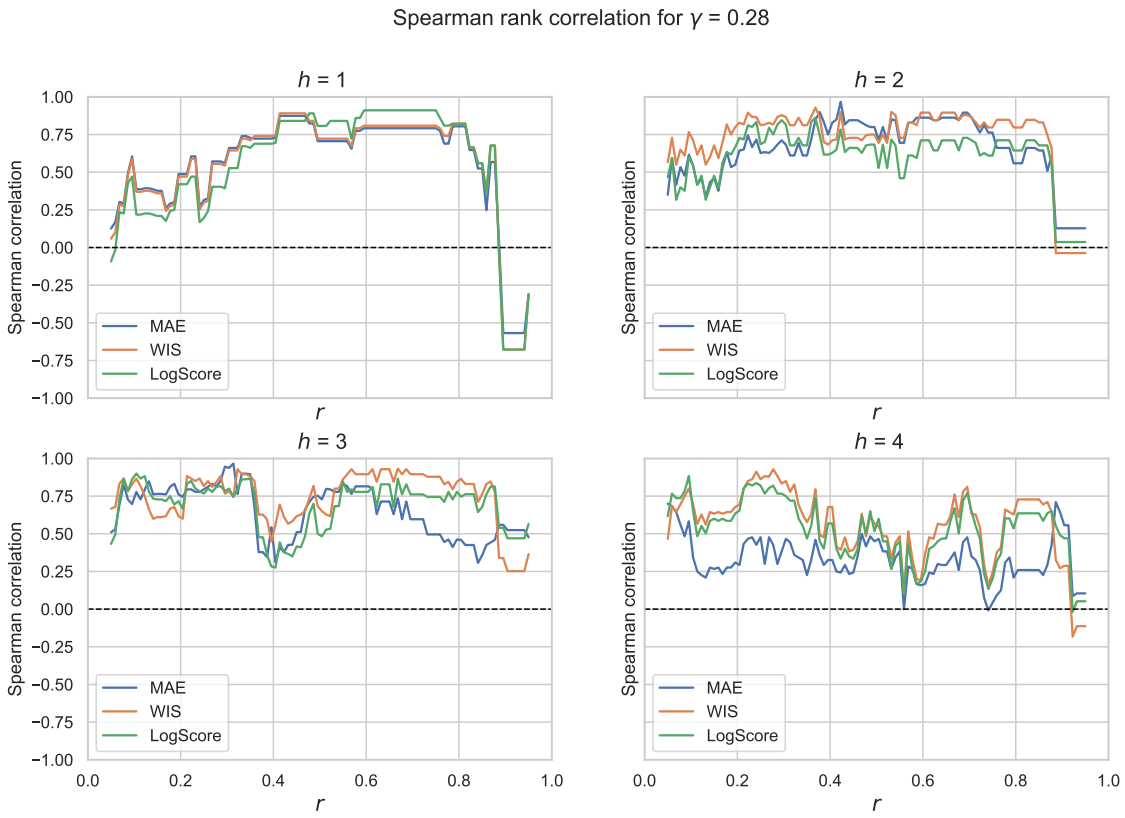

Figure A.11: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.28$.

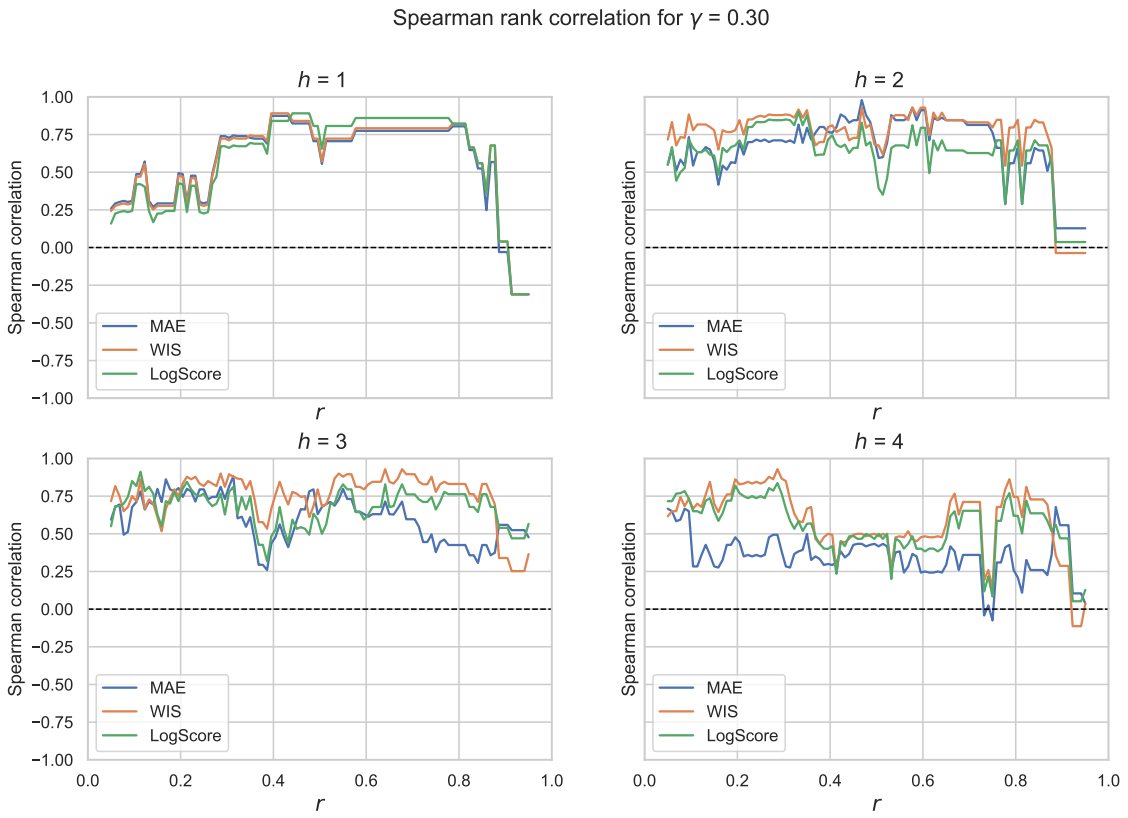

Figure A.12: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.30$.

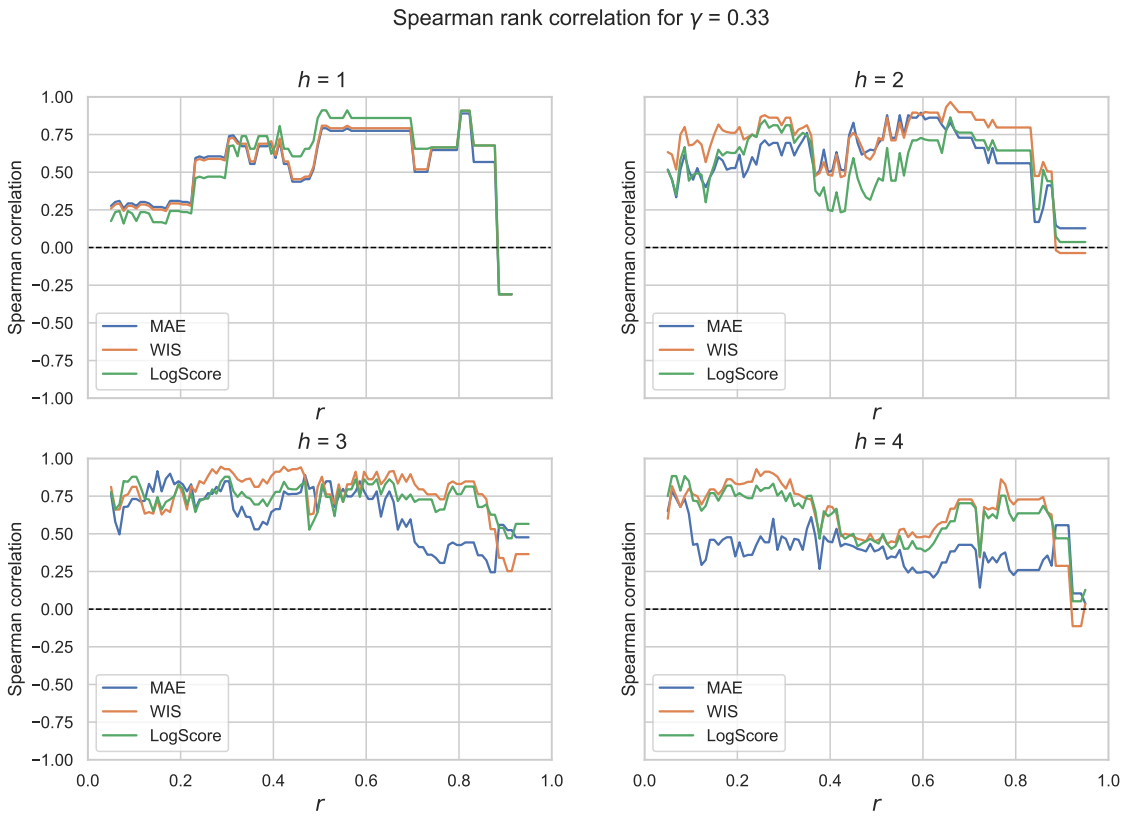

Figure A.13: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.33$.

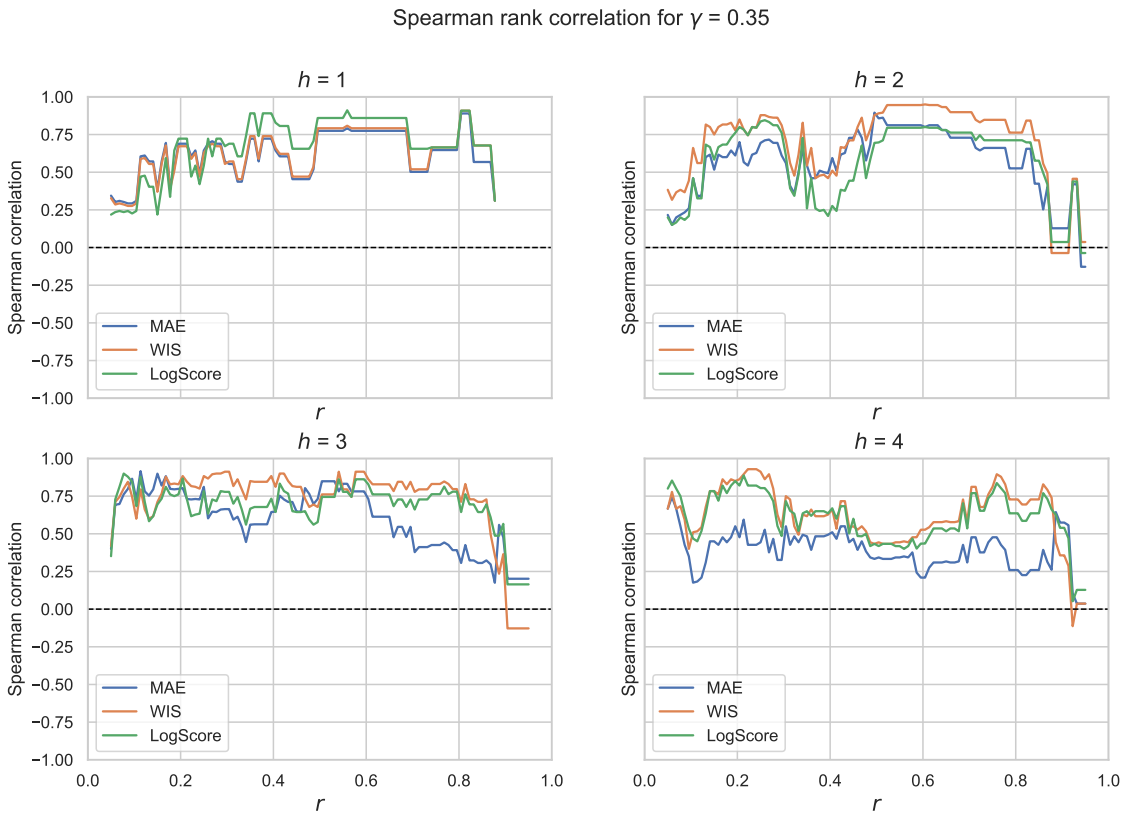

Figure A.14: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.35$.

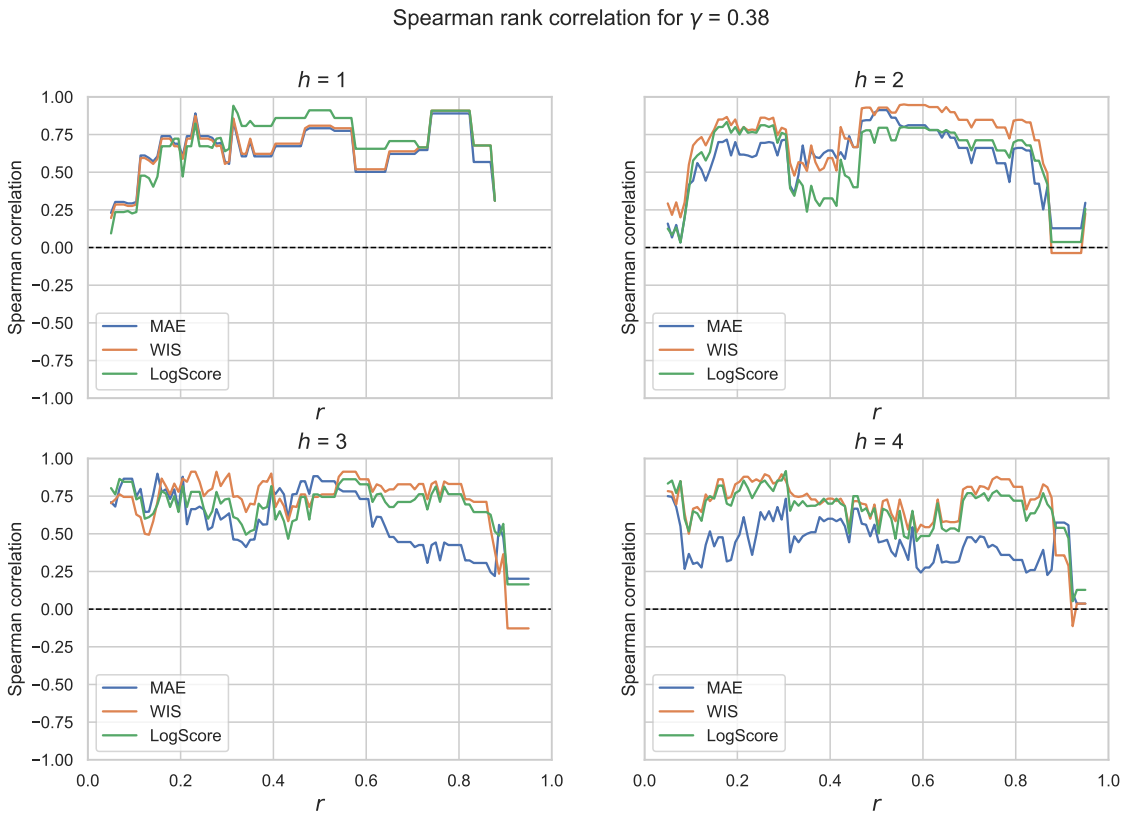

Figure A.15: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.38$.

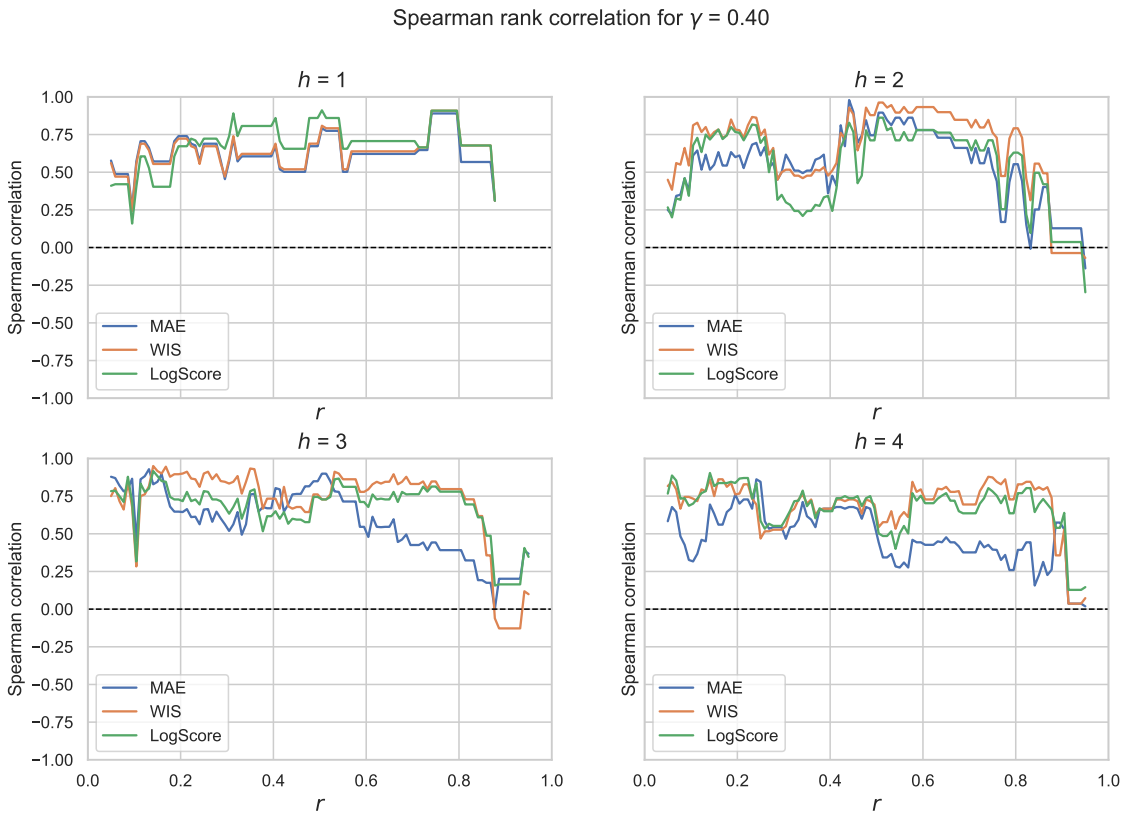

Figure A.16: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.40$.

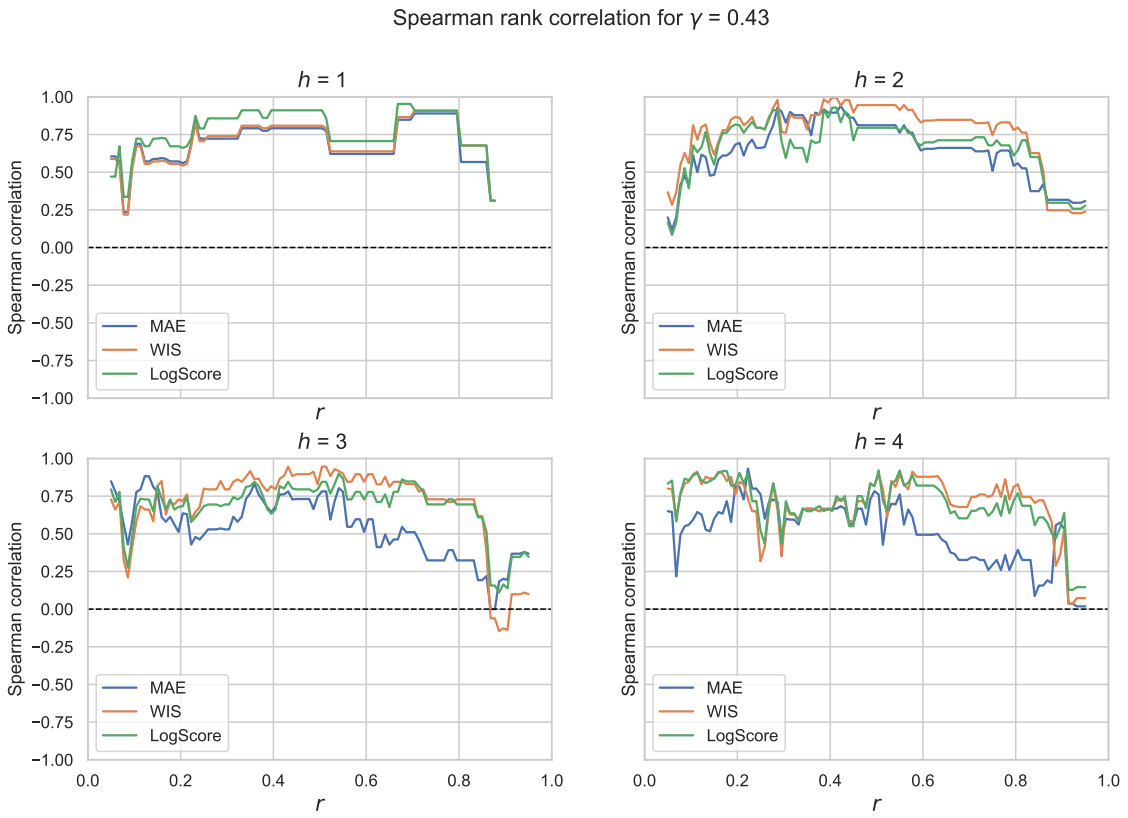

Figure A.17: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.43$.

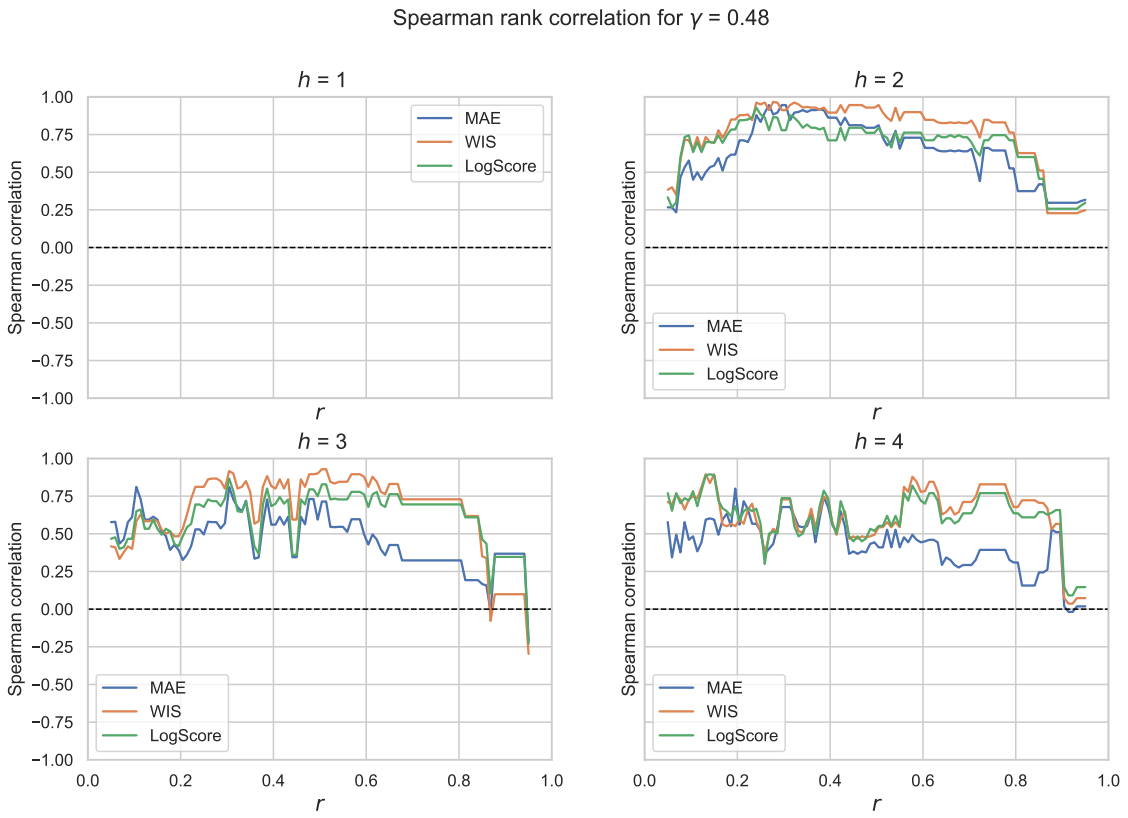

Figure A.18: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.48$.

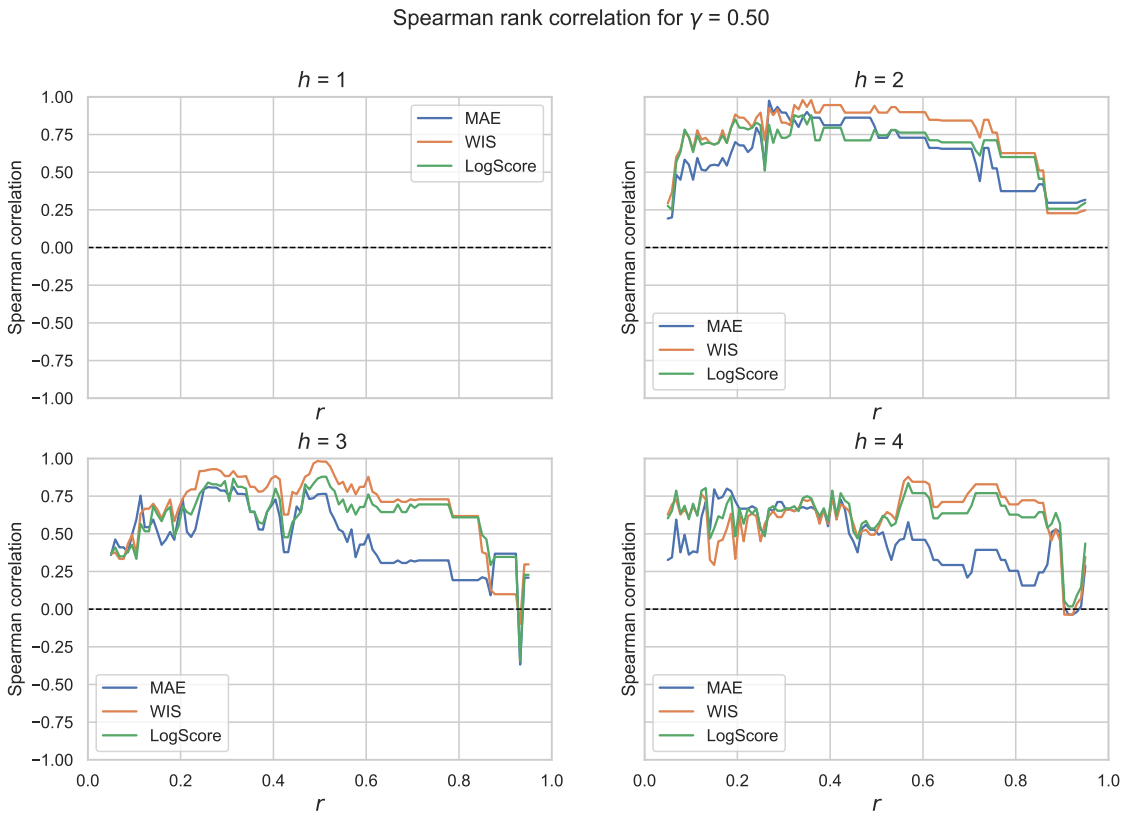

Figure A.19: Spearman rank correlation between forecast accuracy rankings and value score rankings for event definition $\gamma = 0.50$.

B

Appendix 2

This appendix presents additional examples of the cost–loss decision framework corresponding to the remaining outcomes in Table 2.1. Similar to Figure 3.1, each example illustrates how a probabilistic forecast is translated into a decision through the cost–loss framework and the resulting outcome.

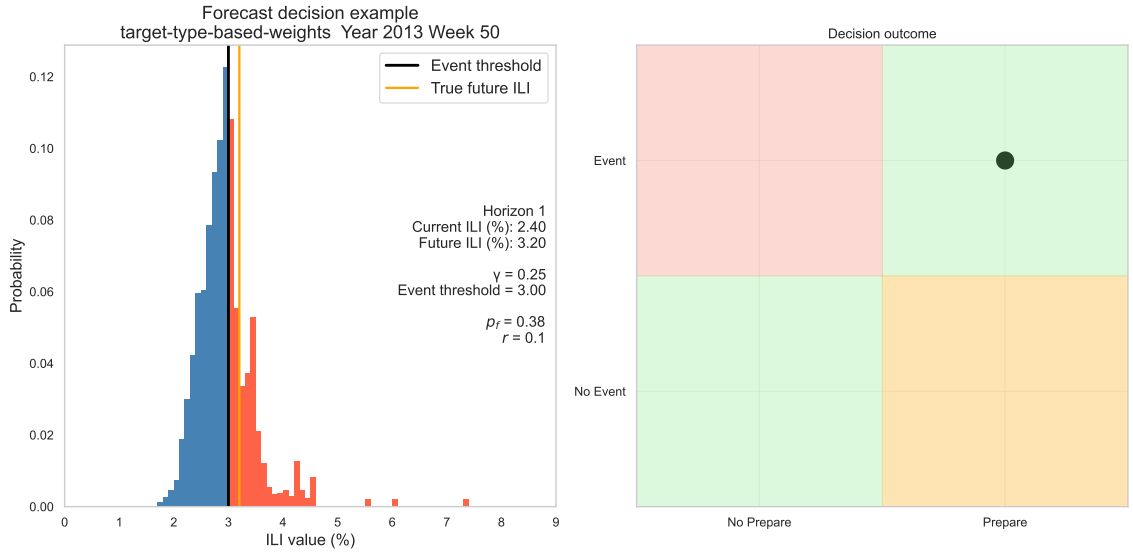

Figure B.1: Example of the cost–loss decision framework where preventive action is taken, and the event occurs. The left panel shows the probabilistic forecast together with the event threshold used to compute the forecast probability p_f, while the right panel displays the corresponding decision outcome in the cost–loss matrix.

In this example, the forecast probability exceeds the cost–loss ratio threshold, and action is taken. Since the event subsequently occurs, the decision results in a cost of C.

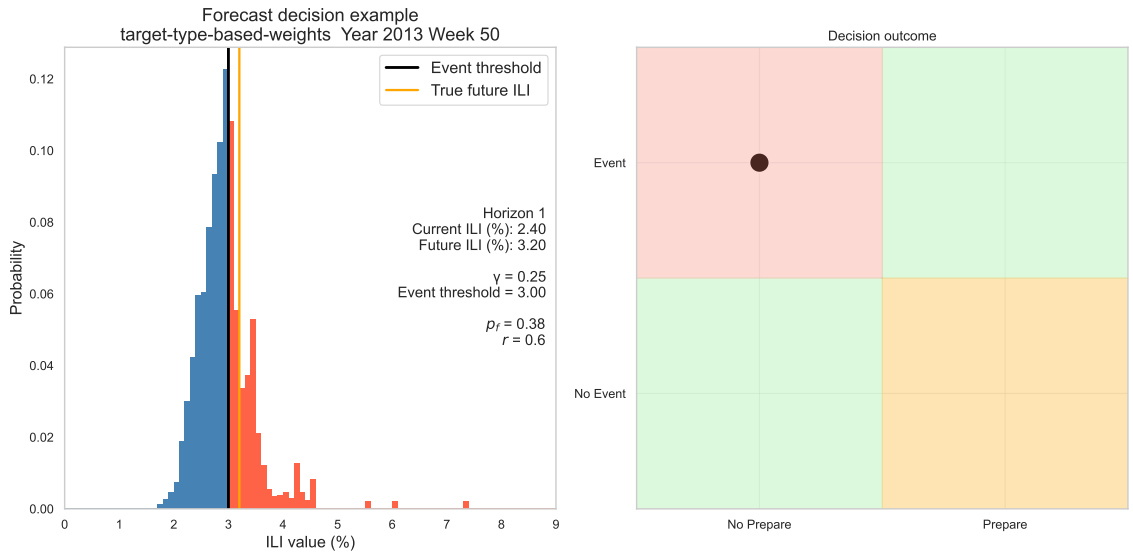

Figure B.2: Example of the cost–loss decision framework where no preventive action is taken, and the event occurs.

In this example, the forecast probability does not exceed the decision threshold, and therefore, no action is taken. Since the event occurs, a loss of L is incurred.

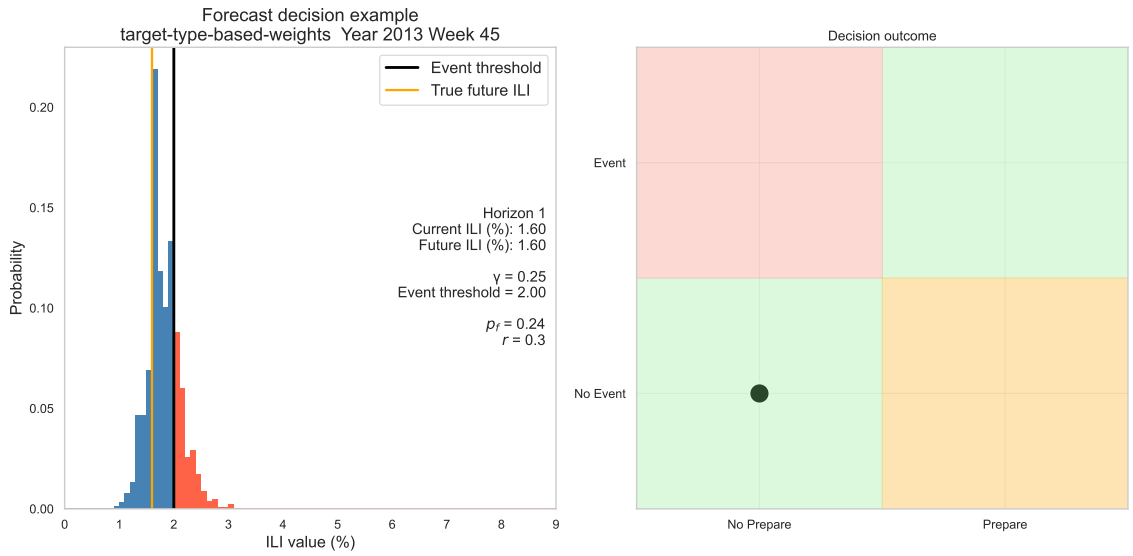

Figure B.3: Example of the cost–loss decision framework where no preventive action is taken, and the event does not occur.

In this example, the forecast probability remains below the decision threshold, and no action is taken. Since the event also does not occur, neither cost nor loss is incurred.

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY