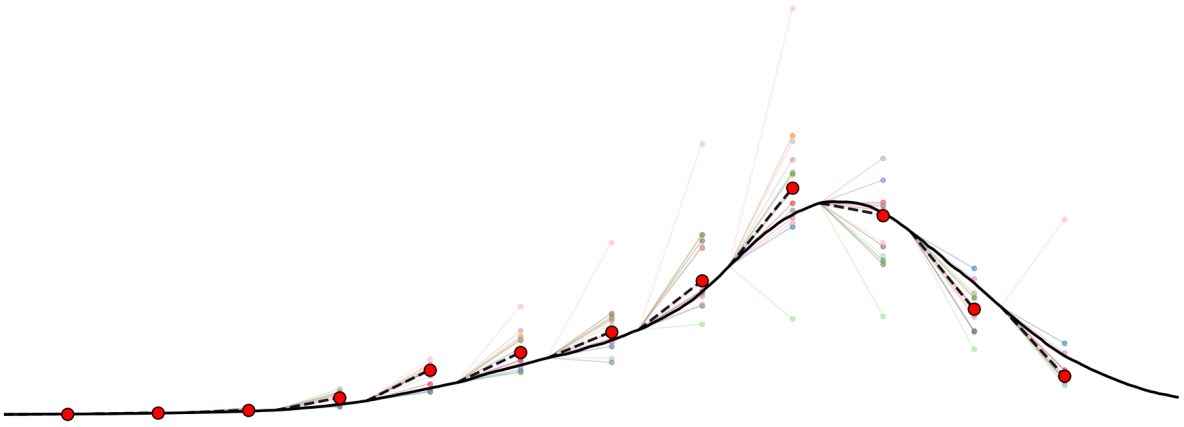


# Infectious disease forecasting using multi-model ensembles

Master's thesis in Engineering Mathematics and Computational Science

Nils Grimbeck

DEPARTMENT OF MATHEMATICAL SCIENCES

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

Master's thesis 2026

# Infectious disease forecasting using multi-model ensembles

Nils Grimbeck


CHALMERS
UNIVERSITY OF TECHNOLOGY

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

Infectious disease forecasting using multi-model ensembles
Nils Grimbeck



Supervisor & Examiner: Philip Gerlee

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: An infectious disease hospitalization time-series with model predictions and the median ensemble aggregation (in red).

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

Infectious Disease Forecasting Using Multi-Model Ensembles
NILS GRIMBECK
Department of Mathematical Sciences
Chalmers University of Technology

## Abstract

Accurate short-term forecasting of infectious disease hospitalisations is critical for healthcare resource allocation and timely public health interventions. Multi-model ensemble forecasts, which combine predictions from several independent models into a single aggregate, have consistently outperformed individual models in real-world forecasting hubs, yet the theoretical reasons for this remain incompletely understood. This thesis develops a probabilistic framework for multi-model ensemble aggregation that makes explicit the role of bias, error variance, model correlation, and outlying components in determining ensemble accuracy. The framework is used to characterise and compare mean, median, and weighted aggregation strategies, with particular attention to their robustness under realistic error structures.

The theoretical predictions are evaluated empirically using two simulated datasets of COVID-like hospitalisation trajectories: a synthetic pandemic dataset of 324 epidemic scenarios generated with CovaSim using Swedish demographic parameters, and a set of 96 counterfactual COVID-19 scenarios for Norway. Fifteen forecasting models spanning statistical, autoregressive, and mechanistic approaches are used as ensemble components.

The results confirm that the unweighted median ensemble consistently produces accurate and stable forecasts, outperforming both individual models and weighted ensemble alternatives on the CovaSim dataset and performing competitively on the Norwegian counterfactuals. This robustness is explained theoretically by the median's bounded sensitivity to outlying components and favourable bias properties, at the cost of a efficiency penalty increasing MSE, a trade-off that favours the median given the heavy-tailed error distributions observed empirically. Weighted ensembles show modest advantages only where component model performance is stable across scenarios, indicating that the median is a robust and reliable aggregation strategy in novel disease settings where no historical performance data is available. Ensemble performance is found to be fundamentally limited by pairwise correlation among component model errors, with diversity and individual model skill acting as complementary rather than substitutable requirements.

Keywords: Infectious disease forecasting, Multi-model ensembles, Ensemble aggregation, COVID-19, Epidemiology, Forecasting, Public Health

## Acknowledgements

First and foremost, I wish to express my gratitude to my supervisor Philip Gerlee. Thank you for taking me under your wings, for the interesting discussions, challenges and encouragement. This work wouldn't have been possible without you and I'm looking forward to our future cooperation.

I would also like to thank Torbjörn Lundh, who together with Philip helped get this project off the ground, and Aila Särkkä, who patiently guided me through my first research experience from which I learned a lot about the research process.

Thanks to the modelling team at Folkehelseinstituttet who welcomed me to Oslo and to the Nordic Pandemic Preparedness Modelling Network, funded by Nordforsk (209400 & 200568), and Birgitte Freiesleben de Blasio for providing the mobility funding. Listening to your methodologically diverse work, always anchored in public health, along with your questions and our discussions about the COVID-19 pandemic, has been both inspiring and greatly beneficial to this work.

This thesis also marks the end of my five years as a student at Chalmers. I am grateful to everyone who has been part of that journey – teachers, fellow students, and all those I had the pleasure of working alongside in the student union, where I not only had a great deal of fun but also grew as a person and colleague. And finally, thank you to all those at Chalmers who allowed me to depart from the usual path and created opportunities from my ideas.

Nils Grimbeck
Oslo 260610

# Contents

# 1 | Introduction

The 2020:s began with the COVID-19 pandemic that paralysed large parts of society and led to worse economic and public health outcomes as nations shut down to "flatten the curve", as not to overwhelm healthcare capacity [1]. Uncertainty about how the pandemic would develop and what effect it would have on healthcare contributed to late interventions and reduced quality of care [2].

This thesis focuses on such respiratory infectious diseases which spread primarily through airborne droplets during close contact. Their transmission dynamics are characterised by rapid exponential growth followed by a decline as the susceptible population is depleted or behaviour changes, producing wave-shaped hospitalisation trajectories [3]. Effective infectious disease outbreak response requires legislators and public health professionals to make timely decisions to reduce disease transmission. During the COVID-19 pandemic surveillance data on the number of cases, hospitalizations and disease-associated mortality among others were used to inform policy [4]. This data provides insight into recent trends but to optimally allocate resources and balance public health outcomes in the population, prediction that anticipate if and when changes occur is critical to inform operational decisions in healthcare and policy interventions [4]. Therefore, forecasts of key epidemiological variables have played an important role in a number of disease outbreaks such as Seasonal Influenza, Ebola, Zika virus, and COVID-19 [5].

Forecasting models use historical data to produce predictions of future developments, called forecasts. The time-series relevant for forecasting depends on the characteristics of the disease. In recent years, diverse sets of data have been used to inform infectious disease forecasting ranging from community surveys to electronic health records, incidence testing, hospitalizations and digital sources such as social media and satellite images [6]. Different data sources have different drawbacks, and while mortality data typically remains stable it is often subject to reporting delays [5]. Furthermore, there is often a delay of several weeks from infection to death, which impacts forecasting efforts. Incidence would thus be a better measure of disease spread that usually don't suffer from similar delays, but it is highly impacted by testing strategies. It has thus been argued that hospital admissions is a useful middle-ground [5] as it is collected routinely and is less sensitive to testing strategies. Furthermore, it can be argued that forecasts of hospitalisations are most valuable for healthcare resource allocation [3].

Prediction accuracy degrades rapidly beyond a month due to the non-linear nature of disease transmission and rapid changes in transmission dynamics due to e.g. mutations, interventions, or voluntary social distancing [5]. It is thus generally acknowledged that predictions are typically short-term whereas medium- to long-term scenario projections based on transmission dynamics assumptions can be utilised to explore events probable

to occur or worst-case scenarios conditioned on the modelling assumptions [7]. In this project, we are therefore interested in short-term forecasting, producing predictions 7 and 14 days ahead, horizons chosen to match the operational planning cycles relevant for hospital capacity management.

All models are inherently uncertain and propagate uncertainty in different ways due to the idealizations, omissions, and simplifications modelling inevitably involves [8]. Different modellers make different choices in how to manage uncertainty and in April 2020 the United States Center for Disease Control and Prevention (CDC) created the US COVID-19 Forecast Hub to collect probabilistic forecasts from academic researchers, government agencies, industry groups, and individuals into a public repository, thus incorporating different modelling approaches and assumptions into a single public data-base [9]. Forecasts from the individual models experienced large variation in accuracy whereas ensemble forecasts, where several models are combined into one prediction showed more consistent accuracy and ranked better than all other models on average [4], which supports prior results that indicate that ensemble forecast provide more accurate and low variance forecasts than the component models on which the ensemble is based [10, 11, 12].

Among the aggregation strategies investigated in real-world forecasting hubs, the simple unweighted median of component model predictions has consistently emerged as the strongest performer on average, outperforming both individual models and more sophisticated weighted combinations [4, 13, 14]. This is a striking finding: despite the apparent simplicity of taking a median, attempts to improve upon it by selecting high-performing models or optimising weights based on past performance have not yielded consistent gains. Why the median is so robust, how model diversity affects ensemble accuracy, and when weighted alternatives might be expected to succeed, remain open questions that this thesis sets out to address.

## 1.1 Aims

Three open questions motivate this work. First, why does the median ensemble perform so consistently well compared to both individual models and other aggregation strategies across diverse infectious disease time-series? Second, how does similarity in component model structure, and the correlation it induces among model errors, affect ensemble performance? Third, under what conditions, if any, can trained weighted ensembles be expected to outperform the simple unweighted median?

The aim of this project is therefore threefold. First, to develop a probabilistic theoretical framework for multi-model ensemble aggregation that makes explicit the role of bias, error variance, model correlation, and outlying components in determining ensemble accuracy. Second, to use this framework to characterise and compare the properties of mean, median, and weighted aggregation strategies, with particular attention to their robustness under realistic error structures. Third, to evaluate these theoretical predictions empirically using two simulated datasets of COVID-like hospitalization trajectories: a synthetic pandemic dataset generated with CovaSim and a set of counterfactual COVID-19 scenarios for Norway.

# 2 | Theory

In this chapter, we begin by briefly introducing the mathematics of infectious disease biology before turning our attention to ensembles, beginning in the wisdom of crowds and presenting some examples of how ensembles have been employed in statistical learning before turning our attention to ensembles in forecasting.

## 2.1 Infectious disease biology

Infectious diseases have shaped human history profoundly. From the Black Death of the 14th century, which claimed an estimated 100 million lives across Europe, to the 1918 influenza pandemic that infected one third of the world's population and caused more than 50 million deaths, to the recent COVID-19 pandemic, infectious disease outbreaks have repeatedly exposed the vulnerability of human populations and the critical importance of understanding how diseases spread [3].

Infectious diseases spread when a pathogen passes from an infectious individual to a susceptible one. The route of transmission depends on the nature of the pathogen. Respiratory viruses such as influenza and SARS-CoV-2 spread mainly through airborne droplets released during close contact; waterborne diseases such as cholera spread through contaminated water supplies; and vector-borne diseases such as malaria rely on an intermediate host, typically an insect. Regardless of the route, the fundamental dynamic is the same: a pool of susceptible individuals is gradually depleted as the disease propagates through the population [3].

Two features of a disease are particularly important for understanding how fast it spreads. The first is the contact period $B$, the average time it takes for an infectious individual to come into effective contact with another person. The second is the infectious period $C$, the duration over which an infected individual can transmit the pathogen to others. Together, these two quantities determine how many secondary infections a single case will generate.

The most widely used summary measure of disease transmissibility is the basic reproduction number $R_0$, defined as the expected number of secondary infections generated by a single infectious individual introduced into a completely susceptible population [3]. In terms of the contact and infectious periods, it can be expressed as:

$$R_0 = \frac{C}{B}. \tag{2.1}$$

The interpretation is intuitive: the longer a person remains infectious relative to the time it takes to come into contact with another susceptible individual, the more people that

person will infect on average. The value of $R_0$ governs the trajectory of an outbreak in a straightforward way:

- If $R_0 > 1$, each infectious individual generates more than one new case on average, and the disease will spread through the population.
- If $R_0 = 1$, the number of cases remains stable (endemic equilibrium).
- If $R_0 < 1$, each generation of infection is smaller than the last, and the outbreak will eventually die out.

While the concept of $R_0$ is straightforward, measuring it in practice is not. During an outbreak, estimates depend critically on testing coverage, reporting practices, the choice of mathematical model, and its initial conditions. For COVID-19, published estimates ranged from $R_0 = 2$–$4$ for simple exponential growth models to $R_0 = 4$–$7$ for more detailed compartment models [3]. Rather than a fixed property of a pathogen, $R_0$ is best understood as a quantity that reflects both biology and behaviour: while the infectious period $C$ is largely determined by the pathogen, the contact period $B$ can be changed by human behaviour and public health measures. It is thus useful to consider the effective reproduction number $R_t$ to account for the variations in expected number of additional infections due to changing social patterns, seasonality, or changing infectious disease biology.

## 2.2 Ensembles and the wisdom of crowds

The idea of using an ensemble is to create a prediction by combining a collection of simpler base predictors to obtain a more accurate and stable estimate than any individual predictor can provide. One of the earliest examples of ensembles is the wisdom of crowds, which was first published by the statistician Francis Galton in 1906 [15]. He tasked people with independently guessing the weight of an ox and found that the average of the crowd's guesses was within 1% of the true weight.

Rauhut and Lorenz [16] states this in mathematical terms. Consider a random variable $X$ representing the distribution of estimates from a crowd of diverse people. The mean squared error of $X$ with respect to the truth $y$ is given by

$$\mathrm{MSE}(X) = \mathbb{E}[(X - y)^2] \tag{2.2}$$

and the population bias is

$$\mathrm{Bias}(X) = (\mathbb{E}[X] - y)^2. \tag{2.3}$$

It follows that

$$\mathrm{MSE}(X) = \mathrm{Var}(X) + \mathrm{Bias}(X) \tag{2.4}$$

and that the collective error of the belief of the crowd, aggregated with the empirical mean $\bar{X}_N = \sum_{i=1}^{N} X_i/N$ of $N$ samples is

$$\mathrm{MSE}(\bar{X}_N) = \frac{\mathrm{Var}(X)}{N} + \mathrm{Bias}(X). \tag{2.5}$$

The implication of this is that averaging several predictors reduces the variance term while leaving the bias unchanged providing better estimates than any individual can provide. Many different wisdom of crowd aggregation methods exists that are more or less suitable

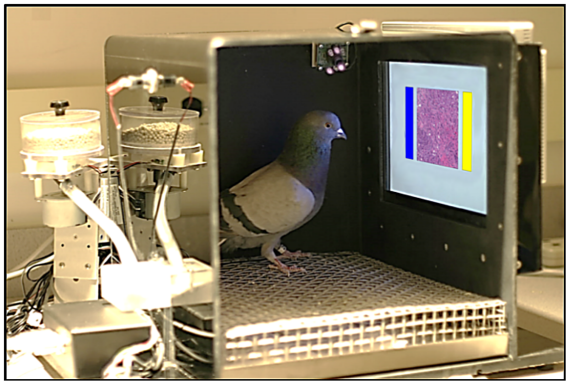


**Figure 2.1:** The pigeons' training environment when distinguishing benign from malignant human breast histopathology. Reprinted under Creative Commons Attribution License from Levenson et al. [17].

for different tasks, but most build on the assumption that crowd guesses or estimates used as input are independent. Deliberation in groups risks enhancing cognitive errors such as bias against the minority, anchoring effects where what is first said affects the outcome, or the common knowledge effect where information held by all members have more influence on the final decision than information held by only a few [15]. Hence, crowds do generally not produce as "wise" estimates as one is lead to believe by equation 2.5 if special care is not taken to ensure independency between predictors.

Levenson et al. [17] provides a commonly used example of the wisdom of crowds in classification tasks. They trained pigeons to distinguish benign from malignant tumours in breast tissue samples, see figure 2.1. The pigeons were found to quickly learn to distinguish benign from malignant histopathology and had an average of 85% accuracy on unseen test data. The researchers also considered a "flock sourcing", where a majority vote from 4 pigeons was used as a classifier. The flock reached a 99% accuracy illustrating the increased performance obtained from crowds, even at small sizes.

During the last decades ensemble methods have grown in popularity within statistical machine learning and several ensemble based algorithms have become commonplace within data science. Bootstrap Aggregation, hereafter refereed to as bagging, is a family of statistical learning algorithms that can be viewed as an ensemble. These algorithms are constructed by repeatedly fitting a model $f(x)$ to a collection of $B$ bootstrap samples $Z^b, b = 1,...,B$, where each bootstrap sample is constructed by sampling $n$ observations with replacement from the original data set $Z$ of size $n$ and averaging. Let $\hat{f}^b(x)$ be the model prediction trained on the bootstrap sample $Z^b$. The bagging estimate is then defined by

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x), \tag{2.6}$$

which has been shown to drastically reduce the variance of unstable predictors, such as trees, and therefore producing an improved prediction [18]. As the same model is used on all bootstrap samples there exists a positive pairwise correlation $\rho$ between the predictors $f^b$ with variances $\sigma^2$. The variance of the average is hence

$$\mathrm{Var}(\hat{f}_{bag}(x)) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{2.7}$$

which limits the benefit of bagging [18]. The random forest algorithm solves this issue by randomly selecting variables at each node in the tree growing process which reduces the correlation between trees and improves performance compared to bagging [18, 19]. Additionally, Boosting, Stacking and Bayesian model averaging can all be viewed as ensemble methods [18] showcasing their widespread use in statistical learning. See *The Elements of Statistical Learning* by Hastie et al. [18] for a thorough treatment.

As illustrated here, ensembles can be constructed in a variety of ways depending on the context and use case. However, it can be broken down into two tasks: developing an ensemble of base models and combining them to form an aggregate predictor.

### 2.2.1 Ensembles in numerical weather forecasting

In the previous section we introduced some ensemble methods used for classification, regression, and estimation, i.e. statistical learning. In infectious disease forecasting we are however interested in using historical data to make predictions about the future. One such context where ensembles have been used for several years is numerical weather forecasting. This section will give an overview of different ensemble methodologies employed within the field and the statistical framework of such climatological ensembles.

Numerical weather prediction (NWP) aims to forecast the future state of the atmosphere, which is an inherently non-linear chaotic system [20, 21]. Tiny differences in initial conditions may thus lead to vastly different predictions and any model will fail to accurately recreate the atmospheric system [22]. Several different ensemble methodologies have been proposed and employed at operational forecasting centres to improve predictions. Du et al. [22] present a comprehensive taxonomy of ensemble-generation strategies, grouped into two major classes: initial-condition ensembles and model-uncertainty ensembles.

Initial-condition methods form the historical core of ensemble predictions in NWP and was introduced to address the inherent uncertainty in the collection and assimilation of observed data into initial conditions [22]. Even with a perfect model, small errors in the initial conditions can amplify rapidly, producing qualitatively different forecasts after only a few days. Initial-condition ensemble methods therefore aim to span the set of initial states that are both meteorologically plausible and capable of producing the range of outcomes possible through the atmosphere's intrinsic instability. Early approaches treated this uncertainty statistically, injecting random perturbations based on climatological error estimates through Monte Carlo methods. While simple, these methods lacked sensitivity to the evolving situation and several methods have been proposed to use dynamically introduced perturbations to reflect the actual uncertainty of the forecast [22].

Initial condition ensembles can reduce random error but not systemic error due to all ensemble members consisting of the same model with different initializations. Model-uncertainty methods aim to address the intrinsic structural error that stems from the mathematical simplifications, limitations of resolution in numerical calculations, and the imperfect treatment of physical processes. Multi-model ensembles is a widely accepted and used method to reduce systemic errors through ensemble averaging and several methods exist to further vary model physics by altering model parameters [22], similar to sensitivity analyses.

### 2.2.2 Reliability in ensemble forecasting

Gneiting et al. [23] proposed that the fundamental goal of a forecast is to maximize sharpness while maintaining calibration/reliability by introducing the following theoretical framework. At a specific time nature chooses a distribution $G_t$ as the true data generating process and a forecaster picks a prediction generated with a predictive distribution $F_t$. An ideal forecast will be drawn from $F_t = G_t$ [23, 24]. An ensemble is thus said to be completely calibrated or reliable if for an $n$-member ensemble $x_{1,t},...,x_{n,t}$ drawn from $F_t$ the true state $y_t$ is drawn from $G_t = F_t \ \forall t$ [24]. While theoretically useful, complete calibration is not testable since $G_t$ cannot be inferred from a single observation $y_t$ at time $t$.

Bröcker & Kantz [25] framed this in terms of exchangeability of the joint distribution of the sequence $Y_t, X_{1,t},...,X_{N,t}$. In this context *"exchangeability implies that the marginal/climatological distribution of each ensemble member equals that of the truth and that potential predictability within the ensemble matches actual predictability relative to the truth"* [24].

A spread-error relationship follows directly from the exchangeability assumption. Consider an $n$-member ensemble $x_{1,t}, \ldots, x_{n,t}$ and a verifying observation $y_t$. Under the exchangeability assumption, the observation is statistically indistinguishable from any ensemble member, meaning that substituting $y_t$ for any member $x_{i,t}$ leaves the joint distribution unchanged [26, 27]. It follows that the mean squared error of the ensemble mean and the ensemble variance must be related. For large ensembles, this reduces to the condition that the root mean squared error of the ensemble mean should equal the square root of the time-averaged ensemble variance:

$$\text{RMSE} \approx \sqrt{\frac{1}{T}\sum_{t=1}^{T} s_t^2} \tag{2.8}$$

where $s_t^2$ is the ensemble variance at time $t$ [26]. An ensemble whose spread is smaller than its RMSE is said to be underdispersed while one whose spread exceeds its RMSE is overdispersed.

Complete calibration also results in a flat rank histogram which was first shown by [23] and follows from the exchangeability assumption [27, 28]. For each forecast, the observation $y_t$ is ranked among the $n$ ensemble members; under complete exchangeability, $y_t$ is equally likely to occupy any of the $n + 1$ possible ranks, so that a histogram of these ranks over many forecast cases should be approximately uniform [27, 28]. Deviations from flatness are informative: a U-shaped histogram indicates underdispersion, where the observation falls outside the ensemble range too frequently, a mound-shaped histogram indicates overdispersion, and a histogram skewed to one side indicates unconditional bias [27]. As Bröcker and Kantz [28] show, the flat rank histogram result holds under the weaker assumption of exchangeability alone, without requiring ensemble members to be independent draws from a common distribution as framed by Gneiting et al [23]. However, a flat histogram is a necessary but not sufficient condition for reliability, as it is insensitive to certain types of miscalibration that only manifest in the joint distribution of ensemble members and observations [27].

As the flat rank histogram and the Spread-Error relationship follows from the reliability assumption, these metrics have been considered a gold standard and therefore heavily relied on for diagnosing ensemble prediction systems within NWP, albeit insufficiently [24,

27]. The reliability assumption is furthermore based on what an ideal forecaster would predict, and due to the highly unstable nature of infectious disease outbreaks there is reason to wonder if this can be assumed to hold in infectious disease forecasting ensembles.

### 2.2.3 Ensembles in infectious disease forecasting

During the COVID-19 pandemic, ensemble forecasts of infectious disease incidence combining multiple epidemiological models consistently outperformed individual models [9, 13]. This has also been demonstrated in the context of seasonal influenza where the FluSight Network's multi-model ensemble approaches have consistently outperformed historical baseline forecasts across regions and seasons [10]. This mirrors the logic of the ensembles presented in section 2.2: diverse models with different structures and assumptions produce a more robust aggregate forecast.

The performance of multi-model ensembles in infectious disease forecasting depends critically on the number of contributing component models according to Becker et al. [14] who investigated this using data from the European COVID-19 Forecast Hub. They found that including more models both improved and stabilised aggregate ensemble performance. The largest gains in performance tended to occur when growing the ensemble from two to four models, after which improvements tapered off or became more gradual. Importantly, the variability of ensemble performance also decreased substantially with ensemble size, suggesting that once a moderate number of models are included there is less need to carefully select which component models to include in the ensemble.

Attempts to improve upon the full unweighted median ensemble by selecting for recent best-performing models or by weighting models according to past performance did not yield consistent gains [14]. Selection ensembles tended to both outperform and be outperformed by the median ensemble across different locations and target series. This is consistent with the model performance observed by Béchade et al. [5] on synthetic data, where no single model consistently outperformed the others across different transmission regimes and no ensemble outperformed the median ensemble on average. Furthermore, Becker et al. [14] found no clear relationship between ensemble diversity, whether measured numerically or through qualitative classification of model types, and ensemble performance. Taken together, these findings support the practical recommendation of aggregating a moderate number of models via the unweighted median, rather than investing resources in model selection or weight optimisation. Similar results were obtained from the US COVID-19 Forecast Hub, where the median ensemble was found to be the best performing aggregation method on average across all component models [4].

## 2.3 Limitation to point predictions

Throughout this thesis, ensemble forecasts are evaluated on point forecast accuracy using RMSE. While probabilistic scoring rules such as the Weighted Interval Score (WIS) and the Continuous Ranked Probability Score (CRPS) are commonly used for evaluating distributional forecasts, their use requires that the ensemble itself produces a well-defined predictive distribution. For multi-model ensembles this is non-trivial: combining quantile forecasts from component models into a joint predictive distribution requires additional assumptions about how the component distributions should be aggregated. One natural approach is quantile averaging, where the ensemble quantile at level $\tau$ is obtained as the weighted mean or median of the component quantiles at the same level, as employed by

[13] in US COVID-19 Forecast Hub ensembles. However, this approach does not in general yield a consistent predictive distribution. To see why, consider two component models that strongly disagree: one predicts a large wave of hospitalisations and one predicts a rapid decline. Quantile averaging would produce an ensemble whose 2.5th percentile is the average of the two lower bounds and whose 97.5th percentile is the average of the two upper bounds. The resulting prediction interval may therefore end up in entirely the wrong location; covering a range that neither component model predicts. For example, if one model produces a 95% interval of $[10, 20]$ and another produces $[110, 120]$, quantile averaging yields $[60, 70]$, which is no wider than either component interval yet falls between them, in a region neither model considers plausible. This is counterintuitive: a situation where models disagree should arguably produce a more uncertain aggregate forecast that spans the range of disagreement. An alternative would be to treat the component forecasts as a mixture distribution, placing weight on each component's full predictive distribution rather than averaging quantile by quantile. This does produce wider intervals when models disagree, but the mixture is highly sensitive to outlying component forecasts, as a single model producing an extreme predictive distribution can dominate the tails of the mixture, which introduces its own challenges around calibration. Given these difficulties, the evaluation in this thesis focuses on point forecast accuracy, which permits a clean comparison of ensemble aggregation strategies without requiring additional distributional assumptions.

# 3 | Implementation

In this chapter we present the methodology of the simulation study by introducing the datasets, forecasting models, and ensemble aggregation methods used to evaluate ensemble performance. In section 3.4 we also introduce a framework for ensemble aggregation and thereafter how it will contribute towards the analysis. Lastly we introduce member-by-member calibration, a method commonly used in numerical weather prediction as a means to improve ensemble forecasts.

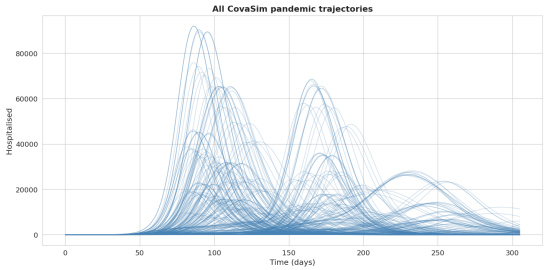
## 3.1 Datasets

To study multi-model ensemble performance across different epidemiological contexts two simulated datasets of COVID-like respiratory infection transmission where used to train and evaluate the forecasting models used in this study, which are introduced in section 3.2.

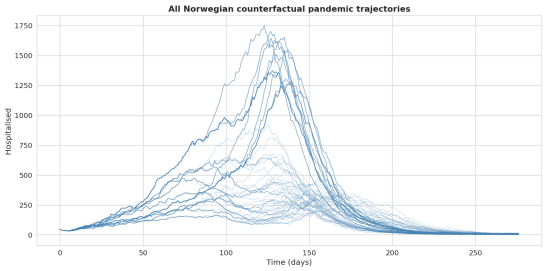
### 3.1.1 CovaSim

A synthetic pandemic dataset generated by Béchade et al. [5] using CovaSim, an agent-based model of COVID-19 transmission, is used to train and evaluate forecasting models in this study. CovaSim simulates the spread of respiratory infectious disease through a population with realistic contact networks across household, school, workplace, and community layers. They configured the model with Swedish demographic parameters, including age distribution and contact network structure, and simulated a population of one million individuals. To produce a maximally diverse set of outbreak trajectories, a Markov Chain Monte Carlo sensitivity analysis was conducted to identify the four parameters with the greatest influence on epidemic shape: the symptomatic-to-severe transition rate, the asymptomatic-to-recovered transition rate, the probability of being symptomatic, and the probability of becoming severe. Each of these four parameters was set to one of three values, half, baseline, or double the default, yielding 81 unique parameter combinations.

To further diversify the dataset beyond disease biology, Béchade et al. [5] applied four distinct time-dependent mobility profiles to each parameter combination as a proxy for varying public health responses and seasonal behaviour: empirical mobility data from Sweden during the COVID-19 pandemic, a sinusoidal seasonal pattern, a step-function representing mandatory lockdowns, and a constant baseline mobility. Combining the 81 disease parameter combinations with these four mobility profiles produced 324 unique epidemic trajectories, each 300 days in length. From each trajectory, daily hospitalisation counts were extracted along with incidence, mobility, and the effective reproduction number.




**(a)** CovaSim trajectories




**(b)** Norwegian counterfactual trajectories

### 3.1.2 Counterfactual COVID-19 scenarios in Norway

To independently validate multi-model ensemble performance a dataset of counterfactual scenarios generated by Diz-Lois Palomares et al. [29] at the Norwegian Institute of Public Health using an agent-based model was also used. The model was originally developed to study methicillin-resistant *Staphylococcus aureus* transmission in Norway and subsequently repurposed to simulate COVID-19 dynamics. The model simulates the spread of SARS-CoV-2 within a synthetic population of approximately 5.4 million agents, replicating the real socio-demography and geography of Norway. Individuals are distributed across a geo-located grid and assigned to households, schools, universities, or workplaces based on their age. The model employs an extended SEIR (Susceptible-Exposed-Infectious-Recovered) framework and explicitly models two co-circulating variants — Delta and Omicron BA.1/BA.2 — along with age-specific infection-hospitalisation ratios, vaccine effectiveness, waning immunity, and non-pharmaceutical interventions.

They calibrated the model to age-specific daily COVID-19 hospital admission data from Norway during the period September 2021 to April 2022, using Bayesian history matching with Gaussian process emulation. Following calibration counterfactual simulations were generated by altering the timing of the booster vaccine rollout that began in late November 2021 as well as a scenario in which no booster doses were administered. The counterfactual dataset used in this study consists of 96 simulated hospitalisation time-series, generated by varying four parameters while keeping all other model parameters fixed. The first is the shift in the timing of the booster vaccine dose rollout, taking four values: $-30$, 0, $+30$, and $+60$ days relative to the baseline, where negative values correspond to an earlier rollout. The second is the adherence to self-isolation recommendations, taking two values: 0.70 (high) and 0.35 (low). The third is the maximum number of vaccine doses an individual can receive, taking two values: 2 and 3. The fourth controls the timing of changes in the transmission rate associated with societal reopening.

The 96 scenarios arise from the full factorial combination of these parameters: 4 booster timing shifts $\times$ 2 adherence levels $\times$ 2 maximum dose caps $\times$ 6 reopening transmission timings, yielding daily hospitalisation time-series that reflect a range of plausible epidemic trajectories under alternative public health conditions. From each scenario, daily hospitalisation counts were extracted along with incidence, mobility, and the effective reproduction number. Compared to the CovaSim data this dataset could be regarded as being more realistic, as it captures the transmission dynamics of two circulating variants in the population as well as more noise. A 7-day rolling average smoothing is thus applied before model training to stabilise forecasts.

## 3.2 Forecasting models

The forecasting models used in this study mainly follow the implementations described in [5], and the reader is referred to that work for full methodological details. A brief description of each model is provided below.

**Exponential regression (ExpReg)** assumes that the number of hospitalised cases follows $h(t) = ae^{b(t-c)}$, with parameters estimated by least squares optimisation. Confidence intervals are computed using the delta method.

**Multivariate exponential regression (ExpMultiReg)** extends the univariate case by incorporating mobility $m_t$ and incidence $i_t$ as additional covariates in the exponent: $h(t) = ae^{(b+dm_t+ei_t)(t-c)}$.

**Moving average (MA)** serves as the baseline model. The point prediction is the mean number of hospitalisations over the preceding seven days, and the confidence interval is given by the corresponding standard deviation.

**ARIMA** fits an ARIMA$(p, d, q)$ model to the hospitalisation time series, with parameters $p = 3$, $d = 0$, $q = 3$, identified via grid search. Estimation and prediction intervals are obtained using the `statsmodels` library.

**Linear regression (LinReg)** is an autoregressive model of order $p = 20$, where past observations serve as covariates and parameters are estimated by ordinary least squares using `scikit-learn`. Confidence intervals are derived analytically from the regression variance.

**Bayesian regression (BayesReg)** is identical in structure to the linear regression model but uses Bayesian ridge regression, placing a Gaussian prior on the weights and estimating hyperparameters via Bayesian inference.

**Vector Autoregression (VAR)** is a multivariate autoregressive model fitted jointly on hospitalisations, incidence and mobility. Parameters are estimated by maximum likelihood using `statsmodels`, which also provides prediction intervals directly.

**SIRH** is an extension of the classical SIR model with a hospitalised compartment $H$, governed by

$$\begin{cases} \dfrac{dS}{dt} = -\beta \dfrac{SI}{N} \\ \dfrac{dI}{dt} = \beta \dfrac{SI}{N} - \gamma_i I - hI \\ \dfrac{dR}{dt} = \gamma_i I + \gamma_h H \\ \dfrac{dH}{dt} = hI - \gamma_h H \end{cases} \tag{3.1}$$

Four variants are considered (SIRH1–SIRH4) in which the recovery rates $\gamma_i$ and $\gamma_h$ are either fixed at their default values ($\gamma_i = 1/8$, $\gamma_h = 1/18$) or treated as free parameters. Parameters are estimated by least squares and confidence intervals via the delta method.

**SIRH-Multi** (SIRH Multi1 and SIRH Multi2) extends the SIRH model by allowing the contact rate to vary with mobility: $\beta_t = a + b \times m_t$. The two variants differ in whether $\gamma_i$ and $\gamma_h$ are free or fixed.

**SEIR-Mob** is an SEIR (Susceptible-Exposed-Infectious-Recovered) model with time-dependent mobility $\beta_t = a + b \times m_t$, adapted from [30]. A fraction $p$ of infected cases are assumed to require hospitalisation after a delay of 21 days.

### 3.2.1 Renewal equation model

In addition to the models of [5], a renewal equation model was implemented and included as a fifteenth component model and as such it is treated in greater detail here. The renewal equation is a statistical model commonly used to model the propagation of infectious diseases by describing how new infections arise from past ones [31]. In this work, the framework has been adapted to model hospitalisations rather than incidence. The renewal equation describes the number of new hospitalisations at time $t$ as

$$H_t = R_t \cdot \Lambda_t, \qquad \Lambda_t = \sum_{s=1}^{T_{\max}} w_s H_{t-s}, \tag{3.2}$$

where $\Lambda_t$ is the force of infection computed from past hospitalisations, $w_s$ is the probability that the time between a hospitalisation and subsequent hospitalisation arising from that patient equals $s$ days, and $R_t$ is the effective hospitalisation reproduction number, ie. how many new hospitalisations a hospitalisation leads to.

**Estimation of $R_t$.** The effective reproduction number is estimated using the method of Cori et al. [32], which places a Gamma prior on $R_t$ and updates it over a sliding window of width $\tau = 7$ days using a Poisson likelihood for the observed hospitalisations. This yields a closed-form Gamma posterior at each time step, from which the posterior mean $\hat{R}_t$ is extracted.

**Generation time distribution.** The generation time between subsequent hospitalisation are modelled with weights $w_s$ derived from a discretised Gamma distribution. Its shape and scale parameters are fitted to the training data by minimising the sum of squared residuals between the observed hospitalisations and the renewal equation fit.

**Forecasting.** The evolution of $R_t$ is modelled as a log-random-walk and its volatility $\sigma$ is estimated as the standard deviation of the differences of $\log \hat{R}_t$ over the training period. Point forecasts are obtained by simulating a single realisation of the log-random walk

$$\log R_{t+1} = \log R_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2), \tag{3.3}$$

and propagating the renewal equation forward along that path. Prediction intervals are derived analytically without Monte Carlo simulation to reduce computational complexity: at horizon $h$, the distribution of $R_{T+h}$ is log-normal with variance $h\sigma^2$ growing linearly with the forecast horizon. This uncertainty is propagated through the renewal equation, and quantiles of the predicted hospitalisations $H_{T+h}$ are read off in closed form from the resulting log-normal approximation.

### 3.2.2 Testing for reliability

To assess whether model would form a reliable ensemble, a chi-square flatness test is applied to the rank histogram as described in [27]. Under the null hypothesis of exchangeability, i.e. that the outcome is statistically indistinguishable from any of the forecasts, each of the $N$ rank bins should be occupied with equal probability $1/N$, yielding a flat

histogram. The Pearson chi-square statistic

$$\chi^2 = \sum_{k=1}^{N} \frac{(O_k - E_k)^2}{E_k}, \qquad E_k = \frac{N}{N}, \tag{3.4}$$

where $O_k$ is the observed count in rank bin $k$, $E_k$ is the expected count under uniformity, and $N$ is the total number of forecast cases, is compared against a $\chi^2$ distribution with $K-1$ degrees of freedom. A $p$-value below the significance threshold of $\alpha = 0.05$ leads to rejection of the null hypothesis, indicating that the outcome is not exchangeable with the model ensemble and thus that the ensemble is not reliable.

## 3.3 Ensembles

The ensemble methods used in this study follow those described in Béchade et al. [5], with the exception of the rank-based ensemble which has been simplified to use overall ranks rather than transmission-regime-specific ranks. A brief description of each ensemble is provided below. All ensemble weights were obtained by splitting the epidemic scenarios into a training and evaluation set, with each scenario assigned to the training set with probability 0.5, and calibrating the weights on the training set only.

**Mean Ensemble (EnsMean)** takes an unweighted mean of all component model point predictions.

**Median Ensemble (EnsMedian)** uses the median of all component model point predictions.

**Regression Ensemble (EnsReg)** takes a weighted mean of the component model point predictions plus an intercept,

$$Y_t^E = w_0 + \sum_{k=1}^{K} w_k \hat{Y}_t^k, \tag{3.5}$$

where the weights $w_k$ and intercept $w_0$ are obtained by linear regression with the observed hospitalisations as the target variable.

**RMSE-optimised Ensemble (EnsRMSE)** is similar to EnsReg but constrains the weights to sum to unity and excludes the intercept. The weights are obtained by minimising the total RMSE of the ensemble forecast over the training set subject to $\sum_k w_k = 1$.

**Rank-based Ensemble (EnsRank)** selects the five component models with the lowest overall average normalised rank on the training set, regardless of the current transmission regime. Their weights are set to the inverse of their rank and then normalised to sum to unity. The ensemble forecast is thus

$$Y_t^R = \sum_{k=1}^{5} w_k \hat{Y}_t^k, \tag{3.6}$$

where the sum runs over the five top-ranked component models.

## 3.4 A framework for ensemble aggregation

Building on the wisdom-of-crowds framing of Section 2.2, we now develop a probabilistic model that makes the structure of ensemble errors explicit and permits a theoretical comparison of different aggregation strategies.

Assume the outcome is uncertain and can be modelled as $Y \sim N(\mu,\sigma^2)$. We model each of the $N$ ensemble members as a biased signal

$$X_i = Y + b_i + \varepsilon_i, \qquad i = 1,\dots,N, \tag{3.7}$$

where $b_i$ is a (possibly time- or context-dependent) bias and $\varepsilon_i \sim N(0,\sigma_\varepsilon^2)$ is a noise term, assumed i.i.d. across members for simplicity. Because every $X_i$ contains the same realisation of $Y$, the predictions are *not* independent; they share a common stochastic component of variance $\sigma^2$ relating to the outcome uncertainty.

### 3.4.1 Mean aggregation

The mean ensemble is given by $\bar{X}_N = N^{-1}\sum_{i=1}^N X_i$ and its expectation and MSE under the signal model 3.7 are

$$\mathbb{E}[\bar{X}_N] = \mu + \bar{b}, \qquad \bar{b} = \frac{1}{N}\sum_{i=1}^N b_i, \tag{3.8}$$

$$\mathrm{MSE}(\bar{X}_N) = \sigma^2 + \frac{\sigma_\varepsilon^2}{N} + \bar{b}^2. \tag{3.9}$$

The term $\sigma^2$ is irreducible and it reflects genuine outcome uncertainty shared by every member. The term $\sigma_\varepsilon^2/N$ decreases monotonically with ensemble size, reproducing the classic variance-reduction benefit of averaging common to many ensemble methods as shown in chapter 2. The squared mean-bias $\bar{b}^2$ is unaffected by $N$ and represents the residual bias.

### 3.4.2 Median aggregation

Let $Z_i = b_i + \varepsilon_i$ denote the total error of member $i$, so that $X_i = Y + Z_i$. Because the $Z_i$ are independent across members we have that

$$\mathbb{E}[\tilde{X}_N] = \mu + m_Z, \qquad m_Z = \mathbb{E}[\mathrm{median}(Z_1,\dots,Z_N)], \tag{3.10}$$

$$\mathrm{MSE}(\tilde{X}_N) = \sigma^2 + \frac{1}{4N f_Z(m_Z)^2} + m_Z^2, \tag{3.11}$$

where $f_Z$ is the density of $Z_i$. To obtain a more interpretable expression, we use the approximation

$$\mathbb{E}[\mathrm{median}(Z_1,\dots,Z_N)] \approx \mathbb{E}[\mathrm{median}(b_1,\dots,b_N)] \tag{3.12}$$

This approximation is valid when (i) the noise $\varepsilon_i \sim N(0,\sigma_\varepsilon^2)$ is symmetric around zero, which ensures $\mathbb{E}[\mathrm{median}(\varepsilon_i)] = 0$, and (ii) $\sigma_\varepsilon$ is small relative to the spread of the $b_i$, so that the rank ordering of the $Z_i$ closely tracks that of the $b_i$. Under these conditions the variance of the sample median of $Z_1,\dots,Z_N$ can be approximated by the variance of the sample median of $N$ i.i.d. $N(0,\sigma_\varepsilon^2)$ variables,

$$\mathrm{Var}(\mathrm{median}(Z_1,\dots,Z_N)) \approx \frac{1}{4N\, f_\varepsilon(0)^2} = \frac{\pi\,\sigma_\varepsilon^2}{2N}, \tag{3.13}$$

where $f_\varepsilon(0) = (2\pi\sigma_\varepsilon^2)^{-1/2}$ is the density of $\varepsilon_i$ evaluated at zero. Intuitively, when the $b_i$ are well-separated relative to $\sigma_\varepsilon$, the identity of the median member is nearly deterministic, and the residual variance comes entirely from the noise $\varepsilon_i$ of that member.

It is worth noting that $\frac{\pi\sigma_\varepsilon^2}{2N}$ is a lower bound on the true variance of the sample median of the $Z_i$. To see this, note that the asymptotic variance of the sample median depends on the density of $Z_i$ evaluated at its median $m_Z$. Approximating the $Z_i$ as i.i.d. draws from a mixture, this density is

$$f_Z(m_Z) = \frac{1}{N}\sum_{i=1}^{N} f_\varepsilon(m_Z - b_i). \tag{3.14}$$

Since $f_\varepsilon$ is the Gaussian density, it is maximised at zero, so $f_\varepsilon(m_Z - b_i) \leq f_\varepsilon(0)$ for all $i$, with equality only when $b_i = m_Z$. Therefore $f_Z(m_Z) \leq f_\varepsilon(0)$, which gives

$$\frac{1}{4N f_Z(m_Z)^2} \geq \frac{1}{4N f_\varepsilon(0)^2} = \frac{\pi\sigma_\varepsilon^2}{2N}. \tag{3.15}$$

The approximation in (3.13) thus systematically underestimates the variance of the median, with the underestimation worsening as the spread of the $b_i$ increases.

Substituting into (3.12) and (3.13) into (3.11) yields

$$\mathrm{MSE}(\tilde{X}_N) \approx \sigma^2 + \frac{\pi\,\sigma_\varepsilon^2}{2N} + \mathrm{median}(b_1, \ldots, b_N)^2. \tag{3.16}$$

Comparing (3.9) and (3.16), the two aggregators are qualitatively similar but differ in two respects. The variance term for the median is $\pi/2 \approx 1.57$ times larger than for the mean, showing an efficiency loss of median aggregation under Gaussian noise, even as the variance term is underestimated. The bias term, however, is governed by the *median* rather than the mean of the $b_i$. The median is more robust to outlying biases, a property that will be exploited in Section 3.4.4.

### 3.4.3 The necessity of diversity

The MSE expressions derived above implicitly assumed that the idiosyncratic errors $\varepsilon_i$ are uncorrelated across members. In practice, ensemble members often share model components, training data, or structural assumptions, introducing positive correlation and shared biases. We now show that such correlation fundamentally limits the benefit of ensemble aggregation for any finite $N$.

#### 3.4.3.1 Correlated mean aggregation

Suppose the errors $\varepsilon_i$ are equicorrelated with pairwise correlation $\rho \in [0,1)$ and common variance $\sigma_\varepsilon^2$. The variance of the mean of $N$ such variables is

$$\begin{aligned}
\operatorname{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i\right) &= \frac{1}{N^2}\operatorname{Var}\left(\sum_{i=1}^{N}\varepsilon_i\right) \\
&= \frac{1}{N^2}\left[\sum_{i=1}^{N}\operatorname{Var}(\varepsilon_i) + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\operatorname{Cov}(\varepsilon_i,\varepsilon_j)\right] \\
&= \frac{1}{N^2}\left[N\sigma_\varepsilon^2 + N(N-1)\,\rho\,\sigma_\varepsilon^2\right] \\
&= \frac{\sigma_\varepsilon^2}{N} + \frac{N-1}{N}\,\rho\,\sigma_\varepsilon^2 \\
&= \frac{\sigma_\varepsilon^2}{N}(1-\rho) + \rho\,\sigma_\varepsilon^2, \qquad (3.17)
\end{aligned}$$

so that under the assumption of correlated ensemble components the MSE of the ensemble mean becomes

$$\operatorname{MSE}(\bar{X}_N) = \sigma^2 + \rho\,\sigma_\varepsilon^2 + \frac{(1-\rho)\,\sigma_\varepsilon^2}{N} + \bar{b}^2. \qquad (3.18)$$

This expression holds for all $N \geq 1$. At $N = 1$ it reduces to $\sigma^2 + \sigma_\varepsilon^2 + b_1^2$, as expected. As $N$ grows, the term $(1-\rho)\sigma_\varepsilon^2/N$ shrinks but the floor $\sigma^2 + \rho\,\sigma_\varepsilon^2 + \bar{b}^2$ remains, limiting the benefit of aggregation. Equation (3.18) is the direct analogue of the bagging variance formula $\rho\sigma^2 + (1-\rho)\sigma^2/B$ noted by Hastie et al. [18]: for any fixed $N$, increasing $\rho$ raises the MSE by $\rho\sigma_\varepsilon^2(1-1/N)$, and the only way to reduce this penalty is to increase the diversity of the ensemble.

#### 3.4.3.2 Correlated median aggregation

For the median under equicorrelated errors, we seek a decomposition of each $\varepsilon_i$ into a part shared across all members and an independent idiosyncratic part. Specifically, we write

$$\varepsilon_i = \sqrt{\rho}\,U + \sqrt{1-\rho}\,\eta_i, \qquad (3.19)$$

where $U \sim N(0,\sigma_\varepsilon^2)$ is a common factor and $\eta_i \overset{\text{iid}}{\sim} N(0,\sigma_\varepsilon^2)$ are independent idiosyncratic components, with $U$ and all $\eta_i$ mutually independent. To verify that this reproduces the correct marginal variance and pairwise covariance, note that

$$\operatorname{Var}(\varepsilon_i) = \rho\,\operatorname{Var}(U) + (1-\rho)\,\operatorname{Var}(\eta_i) = \rho\,\sigma_\varepsilon^2 + (1-\rho)\,\sigma_\varepsilon^2 = \sigma_\varepsilon^2, \qquad (3.20)$$

$$\begin{aligned}
\operatorname{Cov}(\varepsilon_i,\varepsilon_j) &= \operatorname{Cov}(\sqrt{\rho}\,U + \sqrt{1-\rho}\,\eta_i,\ \sqrt{\rho}\,U + \sqrt{1-\rho}\,\eta_j) \\
&= \rho\,\operatorname{Var}(U) + \sqrt{\rho(1-\rho)}\,\operatorname{Cov}(U,\eta_j) + \sqrt{\rho(1-\rho)}\,\operatorname{Cov}(\eta_i,U) + (1-\rho)\,\operatorname{Cov}(\eta_i,\eta_j) \\
&= \rho\,\sigma_\varepsilon^2 + 0 + 0 + 0 = \rho\,\sigma_\varepsilon^2, \qquad i \neq j, \qquad (3.21)
\end{aligned}$$

as required. Conditionally on $U = u$, each $\varepsilon_i = \sqrt{\rho}\,u + \sqrt{1-\rho}\,\eta_i$ is a shift of the independent variable $\sqrt{1-\rho}\,\eta_i$ by the constant $\sqrt{\rho}\,u$. The location equivariance of the

median then gives

$$\text{median}(\varepsilon_1,\ldots,\varepsilon_N)\,\Big|\,\{U=u\} = \sqrt{\rho}\,u + \text{median}(\sqrt{1-\rho}\,\eta_1,\ldots,\sqrt{1-\rho}\,\eta_N) \tag{3.22}$$

$$= \sqrt{\rho}\,u + \sqrt{1-\rho}\,\text{median}(\eta_1,\ldots,\eta_N), \tag{3.23}$$

where location equivariance states that $\text{median}(c+Z_i) = c+\text{median}(Z_i)$ for any constant $c$, and scale equivariance states that $\text{median}(cZ_i) = c\,\text{median}(Z_i)$ for any constant $c>0$. Since $U$ and $(\eta_1,\ldots,\eta_N)$ are independent, we can write (3.23) as

$$M \;:=\; \text{median}(\varepsilon_1,\ldots,\varepsilon_N)\Big|\,\{U=u\} = \sqrt{\rho}\,U + \sqrt{1-\rho}\,\tilde{\eta}_N, \tag{3.24}$$

where $\tilde{\eta}_N := \text{median}(\eta_1,\ldots,\eta_N)$. We now apply the law of total variance,

$$\text{Var}(M) = \mathbb{E}[\text{Var}(M \mid U)] + \text{Var}(\mathbb{E}[M \mid U]).$$

From (3.24), the conditional mean and conditional variance given $U$ are

$$\mathbb{E}[M \mid U] = \sqrt{\rho}\,U + \sqrt{1-\rho}\,\mathbb{E}[\tilde{\eta}_N] = \sqrt{\rho}\,U, \tag{3.25}$$

$$\text{Var}(M \mid U) = (1-\rho)\;\text{Var}(\tilde{\eta}_N) = \frac{\pi(1-\rho)\,\sigma_\varepsilon^2}{2N}, \tag{3.26}$$

where $\mathbb{E}[\tilde{\eta}_N]=0$ by symmetry of $\eta_i$ around zero, and $\text{Var}(\tilde{\eta}_N) = \pi\sigma_\varepsilon^2/(2N)$ follows from (3.13) applied to the $\eta_i \sim N(0,\sigma_\varepsilon^2)$. Substituting (3.25) and (3.26) into the law of total variance,

$$\begin{aligned}\text{Var}(M) &= \mathbb{E}\left[\frac{\pi(1-\rho)\,\sigma_\varepsilon^2}{2N}\right] + \text{Var}(\sqrt{\rho}\,U)\\ &= \frac{\pi(1-\rho)\,\sigma_\varepsilon^2}{2N} + \rho\;\text{Var}(U)\\ &= \frac{\pi(1-\rho)\,\sigma_\varepsilon^2}{2N} + \rho\,\sigma_\varepsilon^2. \end{aligned}\tag{3.27}$$

and therefore

$$\text{MSE}(\tilde{X}_N) \approx \sigma^2 + \rho\,\sigma_\varepsilon^2 + \frac{\pi(1-\rho)\,\sigma_\varepsilon^2}{2N} + \text{median}(b_i)^2, \tag{3.28}$$

under the assumption of correlated ensemble components.

This expression holds for all $N \geq 1$ and mirrors (3.18) structurally. At $N=1$ it gives $\sigma^2+\sigma_\varepsilon^2+b_1^2$ (consistent with the mean). Both the mean and median converge to the same floor $\sigma^2+\rho\sigma_\varepsilon^2$ as $N$ grows. Both strategies are therefore equally affected by correlation in large ensembles, and diversity remains the central lever for improvement regardless of aggregation method.

#### 3.4.3.3 Correlated weighted regression aggregation

A natural extension is to weight the ensemble members unequally, estimating weights by regressing past observations on past predictions. Let $\hat{y} = \sum_{i=1}^N w_i X_i$ where $\mathbf{w} = (w_1,\ldots,w_N)^\top$ is estimated by ordinary least squares on a historical window of $T$ observations. Under the signal model (3.7) with equicorrelated errors, the covariance matrix of the error vector $(\varepsilon_1,\ldots,\varepsilon_N)^\top$ is $\Sigma = \sigma_\varepsilon^2[(1-\rho)I_N + \rho\mathbf{1}\mathbf{1}^\top]$.

For a fixed weight vector $\mathbf{w}$ with $\mathbf{1}^\top \mathbf{w} = 1$, the MSE is

$$\mathrm{MSE}(\hat{y}) = \sigma^2 + \mathbf{w}^\top \Sigma\, \mathbf{w} + (\mathbf{w}^\top \mathbf{b})^2, \tag{3.29}$$

where the variance term evaluates to

$$\mathbf{w}^\top \Sigma\, \mathbf{w} = \sigma_\varepsilon^2 \left[ (1-\rho)\|\mathbf{w}\|^2 + \rho \right]. \tag{3.30}$$

The minimum of (3.30) over all $\mathbf{w}$ with $\mathbf{1}^\top \mathbf{w} = 1$ is achieved by $w_i = 1/N$ for all $i$ (see Appendix A Lemma 1 for a proof), giving $(1-\rho)\sigma_\varepsilon^2/N + \rho\sigma_\varepsilon^2$, exactly the variance of the unweighted mean (3.18). Thus, for fixed $N$ and any $\rho$, the optimal population weights are uniform whenever the biases are equal, and unequal weights can only reduce the bias term $(\mathbf{w}^\top \mathbf{b})^2$ at the cost of increasing the variance term (3.30) above its minimum. Additionally, the estimation of $\mathbf{w}$ on historical data through eg. ordinary least square estimation would further increase the uncertainty in eq. 3.29.

### 3.4.4 Robustness to outlying models

In practice, some ensemble members may be poorly specified or produce faulty predictions. These are not well described by the signal model of Section 3.4. We formalise this by introducing a contaminated-ensemble model in which a fraction $\alpha \in [0,1)$ of models are "outliers" with inflated noise variance.

We partition the $N$ members into a set $\mathcal{G}$ of $N_g = \lfloor (1-\alpha)N \rfloor$ "good" members with noise variance $\sigma_\varepsilon^2$ and a set $\mathcal{O}$ of $N_o = N - N_g$ outlying members with noise variance $\kappa^2\sigma_\varepsilon^2$, $\kappa \gg 1$. For simplicity, set all biases to zero ($b_i = 0$) so that the comparison isolates the effect of variance contamination.

#### 3.4.4.1 Mean aggregation under contamination

For the ensemble mean, the MSE is

$$\begin{aligned}
\mathrm{MSE}(\bar{X}_N) &= \sigma^2 + \frac{1}{N^2}\left[ N_g\, \sigma_\varepsilon^2 + N_o\, \kappa^2 \sigma_\varepsilon^2 \right] \\
&= \sigma^2 + \frac{\sigma_\varepsilon^2}{N}\left[ \frac{N_g}{N} + \frac{N_o}{N}\kappa^2 \right] \\
&= \sigma^2 + \frac{\sigma_\varepsilon^2}{N}\left[ (1-\alpha_N) + \alpha_N \kappa^2 \right],
\end{aligned} \tag{3.31}$$

where $\alpha_N = N_o/N$ is the exact outlier fraction for this $N$. When $\kappa$ is large, even a small fraction $\alpha_N$ of outliers inflates the MSE substantially: the contamination penalty is $\alpha_N(\kappa^2 - 1)\sigma_\varepsilon^2/N$, growing as $O(\alpha\kappa^2/N)$. Since every member enters the aggregation with equal weight $1/N$, the mean offers no protection against a minority of very noisy members.

#### 3.4.4.2 Median aggregation under contamination

The sample median has a breakdown point of $\lfloor (N+1)/2 \rfloor / N$ [33], meaning that the median cannot be moved outside the range of the good members' predictions until $N_o \geq \lfloor (N+1)/2 \rfloor$, i.e. until $\alpha_N \geq 0.5$. For $\alpha_N < 0.5$, however large $\kappa$ is, the sample median remains within the range of the $N_g$ good members' predictions. It will shift somewhat

relative to the median of the good members alone as the outlying predictions grow in number, but it cannot be pulled outside the range of the good members as long as those make up the majority. The MSE of the median is therefore

$$\mathrm{MSE}(\tilde{X}_N) \approx \sigma^2 + \frac{\pi\,\sigma_\varepsilon^2}{2N}, \tag{3.32}$$

where the variance term retains its dependence on $N$ rather than $N_g$, since all $N$ members contribute to the rank ordering that determines the median. However, the variance term remains independent of $\kappa$: the inflated noise of the outlying members affects the spread of their predictions but, as long as $\alpha_N < 0.5$, cannot shift the median outside the range of the good members. Comparing (3.31) and (3.32), the mean degrades in proportion to $\alpha_N\kappa^2$ while the median's MSE is independent of $\kappa$. The advantage of the median over the mean is therefore

$$\mathrm{MSE}(\bar{X}_N) - \mathrm{MSE}(\tilde{X}_N) \approx \frac{\sigma_\varepsilon^2}{N}\Big[(1-\alpha_N) + \alpha_N\kappa^2\Big] - \frac{\pi\,\sigma_\varepsilon^2}{2N}, \tag{3.33}$$

which grows without bound as $\kappa \to \infty$ for any fixed $N$ and $\alpha_N > 0$.

#### 3.4.4.3 Weighted aggregation under contamination

If the weight estimation window includes sufficient observations in which the outlying behaviour is manifested, regression-based weighting can in principle assign near-zero weight to the outlying members. In that case the estimated weighted aggregator approaches the MSE of the good-member mean $\sigma^2 + \sigma_\varepsilon^2/N_g$, which for large $\kappa$ is smaller than (3.31). When the outlying behaviour is intermittent or first manifests during the forecast period, for instance, a model failing on a novel pathogen variant, the historical weights will not reflect the current failure. In that case the regression aggregator assigns positive weight to the outlying members and its MSE approaches (3.31), recovering the same sensitivity to $\kappa$ as the unweighted mean. The latter is probably closer to reality for forecasting of novel infectious diseases due to the few observations to train on and the effect of interventions on transmission dynamics.

### 3.4.5 Robustness to systemic bias

The contamination model introduced above assumed outlying members were extreme purely through inflated noise variance $\kappa^2\sigma_\varepsilon^2$, with biases independent of group membership. We now consider the general case where outlying members may be extreme in both noise variance and systematic bias, with $|b_i^{(o)}| \gg |b_i^{(g)}|$. This represents, for instance, a subset of models trained on unrepresentative data or whose structural assumptions are violated during the forecast period.

Reintroducing biases into (3.31) gives

$$\mathrm{MSE}(\bar{X}_N) = \sigma^2 + \frac{\sigma_\varepsilon^2}{N}\Big[(1-\alpha_N) + \alpha_N\kappa^2\Big] + \Big[(1-\alpha_N)\bar{b}_g + \alpha_N\bar{b}_o\Big]^2, \tag{3.34}$$

where the variance and bias contamination effects are additive. The mean has no mechanism to discount outlying members regardless of whether their extremity arises from inflated noise, large bias, or both: the variance term grows as $O(\alpha_N\kappa^2/N)$ and the squared bias term as $O(\alpha_N^2\bar{b}_o^2)$, each without bound as $\kappa \to \infty$ or $|\bar{b}_o| \to \infty$ for any $\alpha_N > 0$.

For the median, both forms of model failure are governed by the same breakdown-point argument. As long as $\alpha_N < 0.5$, the median cannot be pulled outside the range of the good members regardless of how large $\kappa$ or $|b_i^{(o)}|$ is. The MSE of the median is thus

$$\mathrm{MSE}(\tilde{X}_N) \approx \sigma^2 + \frac{\pi\,\sigma_\varepsilon^2}{2N} + \mathrm{median}(b_1, \ldots, b_N)^2. \tag{3.35}$$

The variance term is independent of $\kappa$ and the bias term cannot be pulled to an extreme value by a minority of severely biased members. Note that bias robustness and variance robustness operate through the same mechanism and therefore cannot be treated as independent protections: a fraction $\alpha_N$ of members that are simultaneously extreme in both bias and noise still counts as a single group of $N_o$ extreme members, and the breakdown condition remains $\alpha_N < 0.5$ regardless of the source of extremity.

Taken together, the median possesses a unified robustness to model failure, whether that failure manifests as inflated noise, systematic bias, or both, while the mean is sensitive to all such failures in proportion to the fraction and severity of the affected models. This robustness comes at the cost of the $\pi/2$ efficiency factor relative to the mean under the homogeneous Gaussian model, a trade-off that will be studied empirically as described in section 3.5.

### 3.4.6 Summary

The analyses in this section together illustrate four complementary principles:

1. outcome uncertainty $\sigma^2$ is an irreducible floor for all aggregators at every ensemble size.

2. diversity is necessary to achieve efficient variance reduction for any finite $N$, with the correlation penalty being $\rho\sigma_\varepsilon^2(1 - 1/N)$ for the mean and $\rho\sigma_\varepsilon^2(1 - \pi/(2N))$ for the median

3. the median cannot be moved outside the range of the good members' predictions as long as $\alpha_N < 0.5$, making its MSE independent of $\kappa$, while the mean degrades proportionally to $\alpha_N\kappa^2/N$ in variance.

4. the same breakdown-point argument applies to systematic bias: the median bias $\widetilde{b}$ cannot be pulled outside the range of the good members' biases as long as $\alpha_N < 0.5$, regardless of how large $|\bar{b}_o|$ is, whereas the mean bias $\bar{b}$ is shifted by every outlying member in proportion to $\alpha_N\bar{b}_o$, with the squared bias penalty $\alpha_N^2\bar{b}_o^2$ growing without bound for any $\alpha_N > 0$.

Table 3.1 collects the MSE expressions derived in this section and the practical effect of these theoretical predictions will be evaluated against simulations.

## 3.5 Empirical analysis of ensemble error structure

The theoretical framework developed in Section 3.4 rests on assumptions about the bias, error distribution and correlation structure of the ensemble members. We now examine these assumptions empirically using simulation data from $N = 15$ models, with the median as the primary aggregator. The analysis proceeds in three parts by estimating biases, the error distributions, and the correlation structure of the models.

**Table 3.1:** MSE expressions for all $N \geq 1$ under the signal model $X_i = Y + b_i + \varepsilon_i$, $Y \sim N(\mu,\sigma^2)$, $\varepsilon_i \sim N(0,\sigma_\varepsilon^2)$. Equicorrelation $\rho$ among $\varepsilon_i$; $N_o = N - N_g$ members having noise variance $\kappa^2\sigma_\varepsilon^2$ with $\alpha_N = N_o/N$. Here $\bar{b} = (1-\alpha_N)\bar{b}_g + \alpha_N\bar{b}_o$ denotes the overall mean bias and $\widetilde{b} = \text{median}(b_1,\ldots,b_N)$ the overall median bias.

| Setting | Mean | Median |
|---|---|---|
| Baseline ($\rho$=0, $\alpha$=0) | $\sigma^2 + \frac{\sigma_\varepsilon^2}{N} + \bar{b}^2$ | $\sigma^2 + \frac{\pi\sigma_\varepsilon^2}{2N} + \widetilde{b}^2$ |
| Correlated ($\rho$>0, $\alpha$=0) | $\sigma^2 + \rho\sigma_\varepsilon^2 + \frac{(1-\rho)\sigma_\varepsilon^2}{N} + \bar{b}^2$ | $\sigma^2 + \rho\sigma_\varepsilon^2 + \frac{\pi(1-\rho)\sigma_\varepsilon^2}{2N} + \widetilde{b}^2$ |
| Contaminated ($\alpha_N$>0, $\kappa \gg 1$) | $\sigma^2 + \frac{\sigma_\varepsilon^2}{N}\left[(1-\alpha_N) + \alpha_N\kappa^2\right] + \bar{b}^2$ | $\sigma^2 + \frac{\pi\sigma_\varepsilon^2}{2N} + \widetilde{b}^2$ |

### 3.5.1 Bias estimation

Under the signal model (3.7), the bias of model $i$ at time point $t$ is $b_{i,t} = \mathbb{E}[X_{i,t} - Y_t]$. The natural estimator of this quantity is the sample mean of the prediction errors, $N^{-1}\sum_t(X_{i,t} - Y_t)$. However, we expect that the prediction errors of several models exhibit heavy tails and occasional catastrophically large values. In the presence of such outlying observations, the sample mean is a sensitive estimator of the central tendency: a single extreme error can substantially distort the estimated bias, pulling it away from the value that characterises typical behaviour. The sample median, by contrast, has a breakdown point of 50% and is therefore unaffected by any minority of outlying observations. We therefore estimate the bias of model $i$ by the sample median of its prediction errors,

$$\hat{b}_i = \text{median}_{t=1,\ldots,T}(X_{i,t} - Y_t), \tag{3.36}$$

which provides a robust estimate of the systematic tendency of each model under typical operating conditions, insensitive to the occasional catastrophic prediction that would otherwise dominate a mean-based estimator.

To investigate whether biases are stationary or depend on the transmission regime, we stratify the predictions by the effective reproduction number $R_t$ into five regimes in the same manner as Béchade et al. [5]: minimal transmission ($R_t < 0.5$), low transmission ($0.5 \leq R_t < 0.8$), stable ($0.8 \leq R_t < 1.2$), high transmission ($1.2 \leq R_t < 3$) and very high transmission ($R_t \geq 3$). Within each regime $s$, the stratum-specific bias is estimated as

$$\hat{b}_{i,s} = \text{median}_{t \in T_s}(X_{i,t} - Y_t), \tag{3.37}$$

where $T_s = \{t : R_t \in s\}$ is the set of time points belonging to regime $s$.

We visualise the results as a heatmap with cell colour indicating the sign and magnitude of $\hat{b}_{i,s}$. This allows immediate identification of models that are systematically biased.

### 3.5.2 Empirical error distributions

To assess whether the Gaussian noise assumption $\varepsilon_i \sim N(0,\sigma_\varepsilon^2)$ is appropriate, and to examine whether the idiosyncratic error variances are homogeneous across models as as-

sumed in Section 3.4, we estimate the empirical error distribution for each model. The bias-corrected residuals are computed as

$$\hat{\varepsilon}_{i,t} = (X_{i,t} - Y_t) - \hat{b}_i, \tag{3.38}$$

where $\hat{b}_i$ is the robust median bias estimate (3.36). Using the median bias in this subtraction ensures that the residuals are centred on the typical prediction error rather than being shifted by occasional extreme observations. For each model $i$ we then produce a kernel density estimate of $\hat{\varepsilon}_{i,t}$ to assess the shape of the empirical error distribution and identify heavy-tailed or skewed models.

### 3.5.3 Model correlation structure

The diversity analysis of Section 3.4.3 showed that the pairwise correlation $\rho$ among model errors is the key determinant of the irreducible MSE floor. We estimate the $N \times N$ empirical correlation matrix of the bias-corrected residuals (3.38),

$$\hat{\rho}_{ij} = \frac{\sum_t \hat{\varepsilon}_{i,t}\hat{\varepsilon}_{j,t}}{\hat{\sigma}_{\varepsilon i}\hat{\sigma}_{\varepsilon j}\, T}, \tag{3.39}$$

and visualise it as a heatmap with hierarchical clustering applied to group models with similar error structure. The equicorrelation assumption used in Section 3.4.3 implies that all off-diagonal entries should be approximately equal; departures from this. for instance, blocks of highly correlated models alongside more independent ones, indicate a factor structure in the errors and suggest that the effective ensemble size is smaller than $N$.

To assess which pairwise correlations are statistically distinguishable from zero, we apply a permutation test to each entry $\hat{\rho}_{ij}$. Under the null hypothesis $H_0\colon \rho_{ij} = 0$, the residual series of models $i$ and $j$ are exchangeable, so shuffling one series destroys any linear dependence while preserving its marginal distribution. For each pair we draw $B = 5000$ random permutations of $\hat{\varepsilon}_{j,t}$, recompute the sample correlation, and obtain a two-sided $p$-value as the fraction of permuted correlations whose absolute value meets or exceeds $|\hat{\rho}_{ij}|$ [34]. Because we test all $\binom{N}{2}$ pairs simultaneously, a Bonferroni correction is applied, so the rejection threshold is $\alpha/\binom{N}{2}$.

To quantify the aggregate diversity of the ensemble we report the mean pairwise correlation

$$\bar{\rho} = \binom{N}{2}^{-1} \sum_{i<j} \hat{\rho}_{ij}. \tag{3.40}$$

The correlation analysis is repeated stratified by transmission regime. A finding that $\hat{\rho}_{ij}$ is substantially higher in particular regimes would imply that the effective ensemble size and hence the benefit of aggregation is regime-dependent, providing an empirical motivation for regime-specific weighting or trimming strategies.

### 3.5.4 The effect of correlated models on ensemble performance

To investigate whether forecast diversity is predictive of ensemble accuracy, we conducted a systematic analysis of sub-ensembles using the median aggregation formed from the full pool of component models. For ensemble sizes ranging from 4 to 9 models, we sampled up to 300 random combinations of component models from all possible subsets of that

size. For each combination the resulting RMSE was calculated across all pandemics and forecast points.

To quantify forecast diversity, we computed the mean pairwise Pearson correlation $\bar{\rho}$ of the raw prediction errors across all pairs of models within each combination, as defined in equation 3.40. The relationship between $\bar{\rho}$ and RMSE was then examined for each ensemble size using an OLS linear regression, and summarised by the Pearson correlation coefficient $r$ between $\bar{\rho}$ and rank. To visualise the best achievable performance at each level of diversity, a lower envelope was constructed by binning ensembles by $\bar{\rho}$ and recording the minimum rank in each bin.

## 3.6 Member-by-member calibration

To improve the calibration of the multi-model ensemble forecasts, we apply a member-by-member postprocessing (MBMP) approach following the reference MBMP method described in Schefzik [35], which they found outperformed the raw ensemble in NWP. MBMP methods retain the rank dependence structure of the raw ensemble by mapping each ensemble member forecast separately to a postprocessed ensemble member forecast using a monotone transformation.

Concretely, for a fixed prediction time, let $\bar{x}$ denote the ensemble mean and $x_i$ the prediction of the $i$-th ensemble member. The raw ensemble forecast $x_1, \ldots, x_N$ is transformed to a postprocessed ensemble forecast $\hat{x}_1, \ldots, \hat{x}_N$ via the monotone mapping

$$\hat{x}_m = \zeta + \eta\bar{x} + \kappa(x_m - \bar{x}), \tag{3.41}$$

where $\zeta$ is a bias correction term, $\eta$ rescales the ensemble mean, and $\kappa$ adjusts the spread. The parameter design is based on two conditions: the climatological variance of the calibrated ensemble members should equal that of the observations, and the overall correlation of the adjusted ensemble members with the unadjusted ensemble mean should equal the correlation of the truth with the unadjusted ensemble mean. These conditions yield

$$\eta = \rho\frac{s_y}{s_{\bar{x}}}, \qquad \kappa = \sqrt{1-\rho^2}\,\frac{s_y}{s_\varepsilon}, \qquad \zeta = \mu_y - \eta\mu_{\bar{x}}, \tag{3.42}$$

where $\rho$ is the Pearson correlation between ensemble means and observations over the training period, $s_y$ and $s_{\bar{x}}$ are the standard deviations of observations and ensemble means respectively, and $s_\varepsilon = \sqrt{\overline{\mathrm{Var}_m(x_m)}}$ is the square root of the mean within-ensemble variance, capturing the raw ensemble spread.

To avoid data leakage, parameters are estimated in a strictly sequential fashion: for prediction $t$ within a given pandemic time series, only the observations and forecasts from time steps $1, \ldots, t-1$ of that same time series are used for training. No information from other pandemic time series is incorporated. A minimum of three training points is required before calibration is applied; predictions made before this threshold use the raw ensemble median directly. The calibrated ensemble median is then used as the point forecast for evaluation.

We are furthermore interested in how more knowledge about similar pandemics might affect the performance of MBMP. To this means each pandemic is regarded as subsequent waves of the same respiratory infection over several years and the MBMP is constructed sequentially over all time-series. In both cases the MBMP forecast is compared to the median ensemble forecast.

# 4 | Results

In this section we present the results of the simulation study. The results are organised as follows: we first evaluate the performance of the individual forecasting models on both datasets, before turning to the performance of the ensemble methods. We then examine the empirical bias and correlation structure of the component models, evaluate the member-by-member calibration approach, and finally investigate the spread-skill relationship. All results are shown for the CovaSim dataset and when the figures show similar results for the Norwegian counterfactuals the same analysis is not repeated in print. The reader is referred to Appendix B where additional figures are printed.

## 4.1 Performance of forecasting models

In this section we present the relative forecasting results of the 15 models used to create the ensembles. We first turn our attention to the CovaSim dataset before comparing the model's performance on the Norwegian counterfactual data which we expect to be more difficult for forecasting, as it contains two peaks compared to the single peak found in the CovaSim dataset.

### 4.1.1 CovaSim

Figure 4.1 shows the distribution of ranks obtained by the 15 models ranked by RMSE, similarly to the presentation by Béchade et al. [5]. Compared to their analysis we have here included an additional model, the Renewal Equation, which performs well with most of its prediction obtaining lower ranks. The figure shows clear trends in model ranks but also that all models are both the best and worst ranking model at some point.

In table 4.1 the ranks are summarized for both 7-day and 14-day forecasts. The ranks have been normalized such that 0 corresponds to the best possible rank and 1 the worst. It is evident that the Moving Average model, that uses the average of the last 7 days as its forecasts, performs worst out of the models whereas VAR, the Renewal Equation, Exponential regression all obtains low ranks in both cases. It is noteworthy that these models are all built on very different assumptions and utilizes the data in different ways: VAR is an autoregressive model utilizing hospitalizations, incidence and mobility, the Renewal equation is a statistical method built on estimating the effective reproduction number $R_t$ of hospitalizations from a hospitalization time-series, and exponential regression is a statistical method modelling exponential growth from a hospitalization time-series.

In figure 4.2 we expand the normalized rank analysis to five transmission regimes defined in section 3.5.1 from which is it evident that model performance is highly affected by

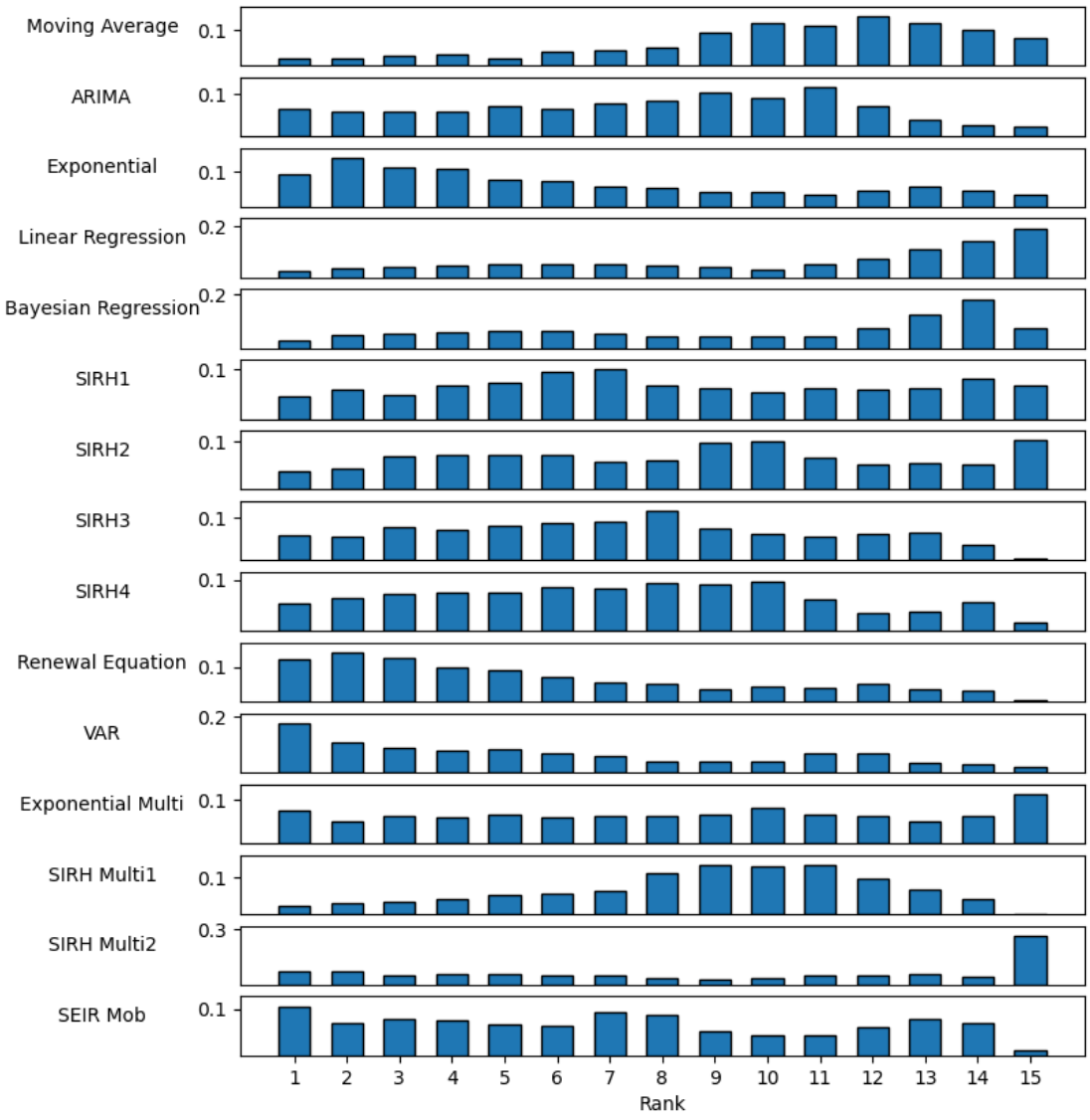



**Figure 4.1:** Histograms of model ranks, ranked by RMSE, for 14-day forecasting on the CovaSim dataset

**Table 4.1:** Average normalized rank for all 15 models, 7-day and 14-day ahead forecasting, evaluated on the CovaSim and the Norwegian counterfactual datasets

| | CovaSim | | Norwegian | |
|---|---|---|---|---|
| Model | 7 days | 14 days | 7 days | 14 days |
| Moving Average | 0.69 | 0.62 | 0.63 | 0.60 |
| ARIMA | 0.38 | 0.44 | 0.39 | 0.43 |
| Exponential | 0.37 | 0.35 | 0.34 | 0.33 |
| Linear Regression | 0.51 | 0.61 | 0.53 | 0.59 |
| Bayesian Regression | 0.45 | 0.56 | 0.50 | 0.53 |
| SIRH1 | 0.51 | 0.48 | 0.47 | 0.47 |
| SIRH2 | 0.54 | 0.50 | 0.36 | 0.35 |
| SIRH3 | 0.44 | 0.42 | 0.58 | 0.58 |
| SIRH4 | 0.46 | 0.43 | 0.40 | 0.38 |
| Renewal Equation | 0.31 | 0.31 | 0.32 | 0.31 |
| VAR | 0.30 | 0.33 | 0.40 | 0.40 |
| Exponential Multi | 0.50 | 0.49 | 0.56 | 0.54 |
| SIRH Multi1 | 0.56 | 0.52 | 0.46 | 0.45 |
| SIRH Multi2 | 0.54 | 0.54 | — | — |
| SEIR Mob | 0.42 | 0.41 | 0.55 | 0.53 |

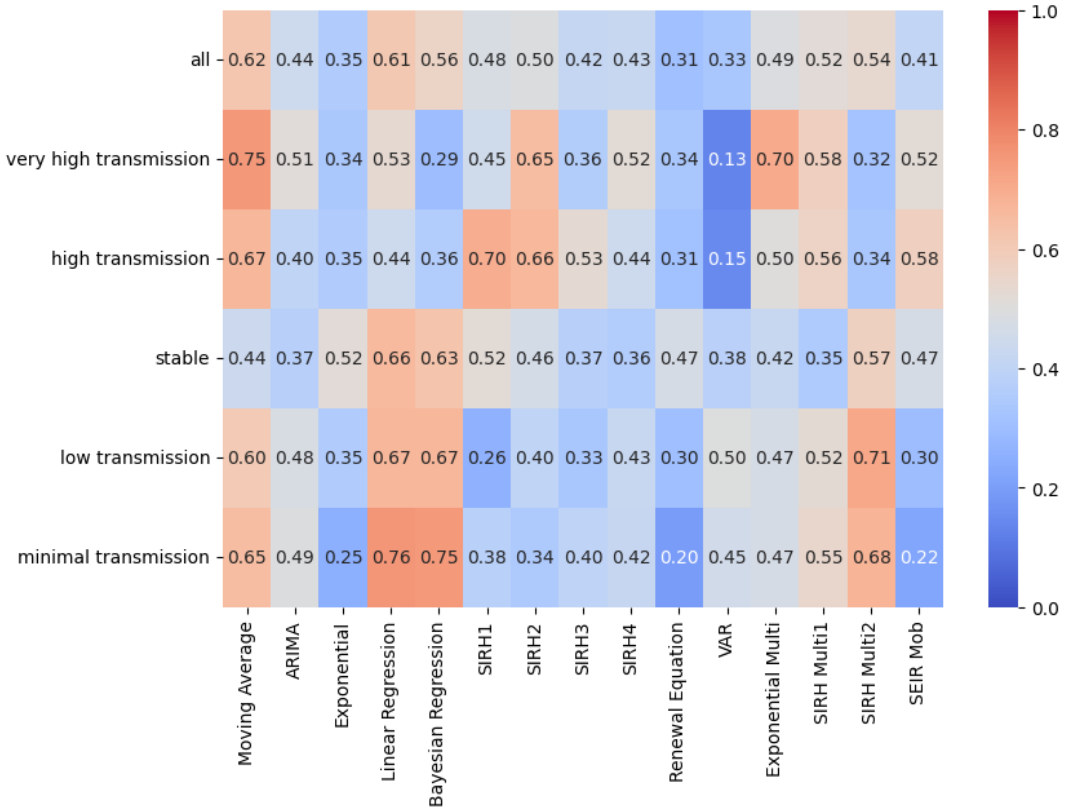



**Figure 4.2:** Average normalized rank of all 15 models and 14-day ahead forecasting, stratified by $R_t$, on the CovaSim dataset

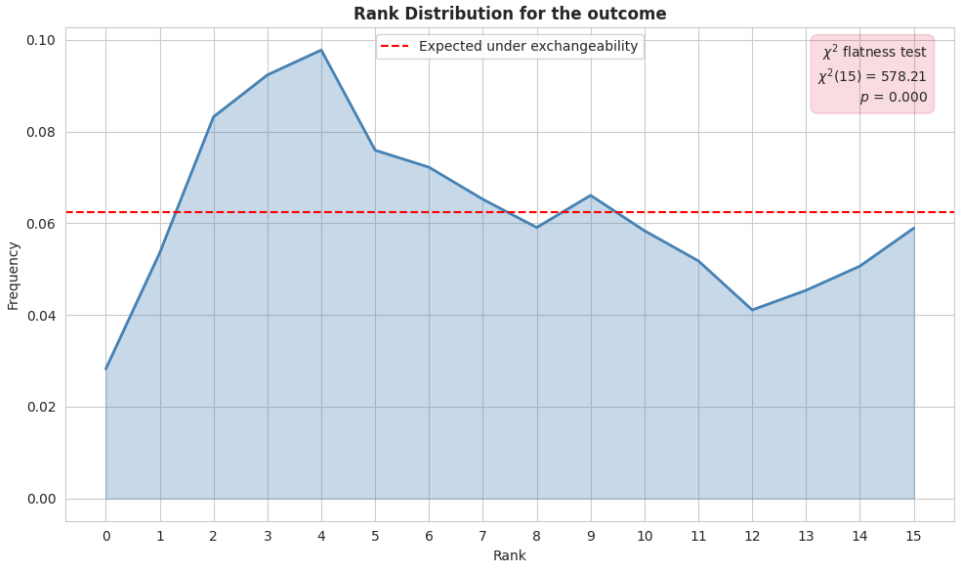



**Figure 4.3:** Rank histogram of the outcome on the CovaSim dataset.

the infectious disease transmission context, further indicating the added benefit of using several forecasts.

We now evaluate whether the ensemble is reliable, under which one would expect a flat rank histogram, where all models and the outcome are ranked from smallest to largest, as described in Section 2.2.2. Figure 4.3 indicates that this is not the case, as the rank histogram for the outcome deviates substantially from the expected uniform distribution, as confirmed by the $\chi^2$ flatness test described in Section 3.2.2, which rejects the null hypothesis of uniformity with $p < 0.001$.

### 4.1.2 Norwegian counterfactuals

The normalized rank results in Table 4.1 and figure 4.4 reveal both consistent trends and notable differences in model performance across the two datasets. Several models perform reliably well on both the CovaSim and Norwegian data: the Renewal Equation is the top or second-best performer across all horizons on both datasets, and simpler statistical models such as Exponential regression and ARIMA occupy the upper-middle tier consistently. Moving Average is the weakest model on both datasets by a clear margin. These consistencies suggest that models which capture the short-term momentum of epidemic dynamics without over-fitting to structural assumptions tend to generalize well regardless of the underlying data generating process.

The more striking differences emerge among the mechanistic models. SIRH2 improves substantially on the Norwegian data (0.36/0.35) relative to CovaSim (0.54/0.50), while SIRH3 moves in the opposite direction, performing reasonably on CovaSim (0.44/0.42) but poorly on Norwegian data (0.58/0.58), suggesting that which parametric assumptions within the SIRH family are beneficial depends strongly on the noise level and epidemic structure of the data. VAR performs best on CovaSim for 7-day forecasting but degrades on the Norwegian data, plausibly because the linear cross-variable dependencies it exploits are cleaner in the CovaSim setting than in the noisier, more realistic Norwegian trajectories. SIRH Multi2 performed poorly on the CovaSim data (0.54/0.54) and was

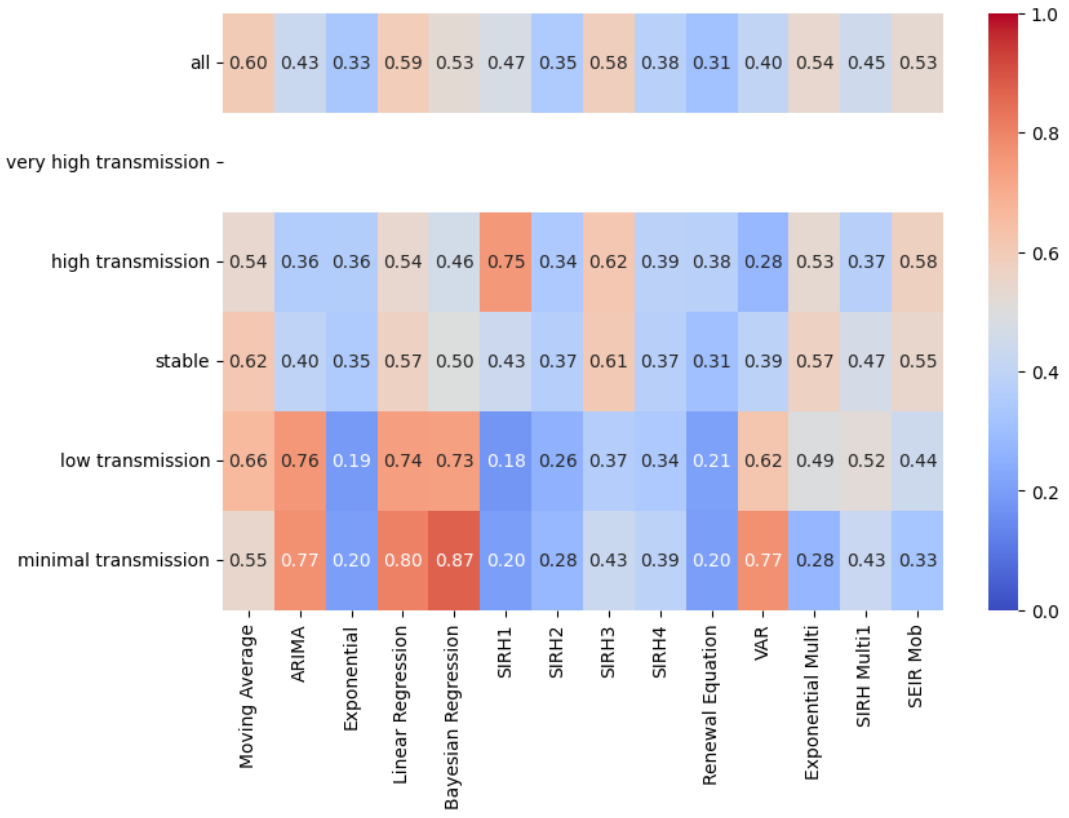



**Figure 4.4:** Average normalized rank of all 15 models and 14-day ahead forecasting, stratified by $R_t$, on the Norwegian counterfactuals dataset

therefore excluded from evaluation on the Norwegian dataset. Overall, the Norwegian dataset, being more realistic by virtue of modelling two co-circulating variants, waning immunity, and greater observational noise, appears to be a more demanding benchmark that differentiates model performance more sharply than the cleaner CovaSim trajectories.

The rank-histogram (not shown) for the Norwegian counterfactuals indicates similarly to the CovaSim dataset that the ensemble is not reliable.

## 4.2 Performance of ensembles

### 4.2.1 CovaSim

The ensembles introduced in Section 3.3 aggregate the predictions of the 15 component models. The estimated weights for each ensemble are shown in Figure 4.5. The three optimised ensembles, EnsReg, EnsRMSE, and EnsRank, assign different weights to the component models, suggesting that the optimal combination of models is sensitive to the choice of loss function.

The average normalized rank for each ensemble when ranked against all models and ensembles is presented in figure 4.6. The median ensemble obtains the lowest rank, as expected by the theoretical derivation in 3.4, but the rank based ensemble where the five best performing models are aggregated through a weighted average according to their ranks is not far behind. Furthermore, ensemble performance is not constant across transmis-

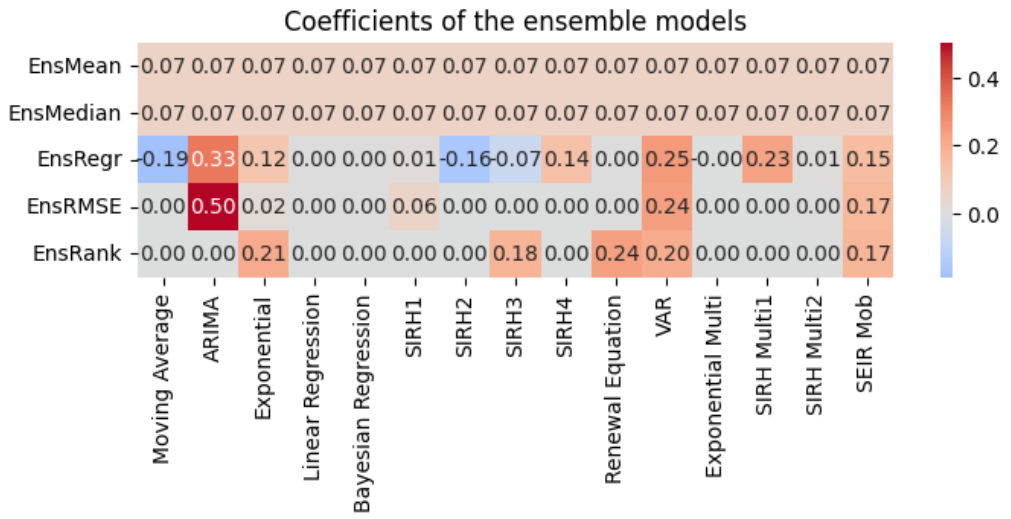



**Figure 4.5:** Coefficients of the ensemble models evaluated on the CovaSim dataset. Model coefficients are fitted on half of the forecasts and ensemble performance is evaluated on the unseen other half.

sion regimes, with the rank based ensemble outperforming the median ensemble in several strata.

Interestingly, the median ensemble and the renewal equation model achieve the same average normalised rank, which might suggest similar prediction accuracy. However, the average normalized rank masks an important difference in the tail behaviour of their error distributions. Figure 4.7 shows a boxplot of the RMSE distributions which reveals that the renewal equation model produces substantially larger errors than the median ensemble, despite their identical average ranks. Plotting RMSE by percentile makes this divergence explicit: up to the 73rd percentile the two models perform comparably, but beyond this point the errors of the renewal equation model grow substantially larger. This indicates that the renewal equation model performs well on the majority of forecasts but fails severely on a minority of cases, resulting in a heavy-tailed error distribution that is obscured by the average rank statistic. Thus, the median ensemble produces more robust forecasts compared to using a single model.

This stability is further illustrated in Figure 4.6, which shows the rank distributions of all ensembles. The mean ensemble is highly skewed towards higher ranks, which is expected under the signal model introduced in Section 3.4 in the presence of contamination and/or systematic biases among the component models. The median ensemble is more stable but is seldom the single best forecast; there is almost always a more accurate prediction contained within the ensemble as evident from figure 4.8. The challenge is therefore to identify which component models will outperform the median ensemble at a given forecast time, but based on the average normalised rank no single model is consistently more accurate. Both EnsReg and EnsRMSE have rank distributions spread across the entire range, offering little improvement over the median in terms of stability. EnsRank more frequently achieves lower ranks than the median ensemble but is also more spread across the rank range.

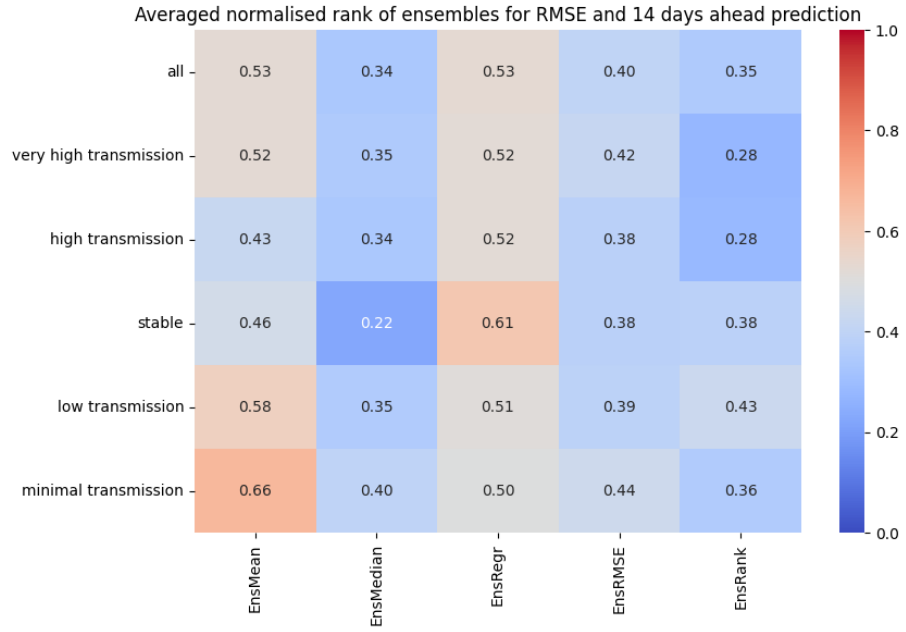



**Figure 4.6:** Average normalized rank of the ensembles across transmission regimes on the CovaSim dataset.

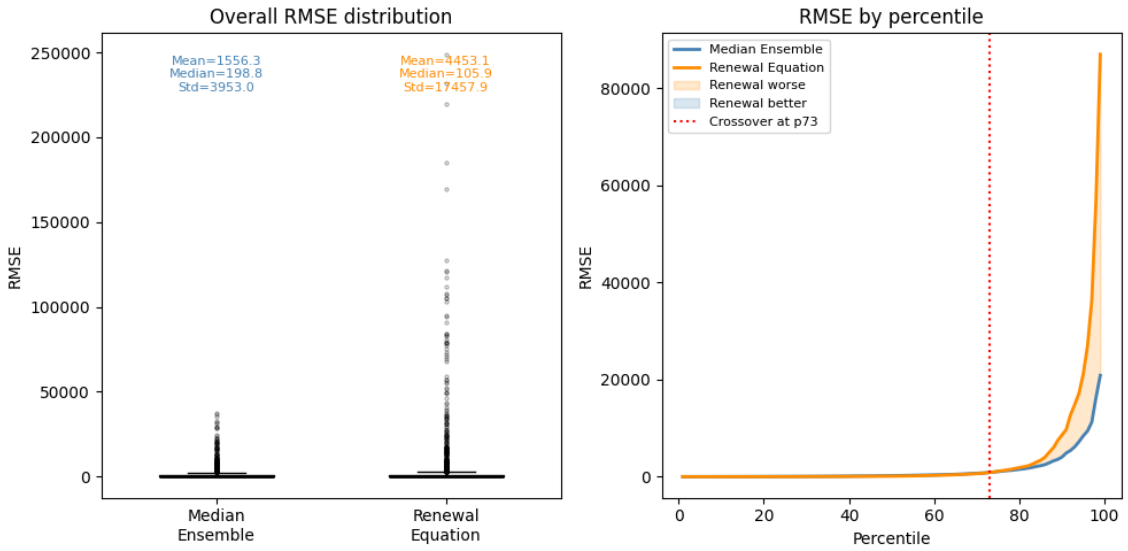



**Figure 4.7:** Comparison of the tail distributions of RMSE for the Renewal Equations and the median ensamble on the CovaSim dataset.

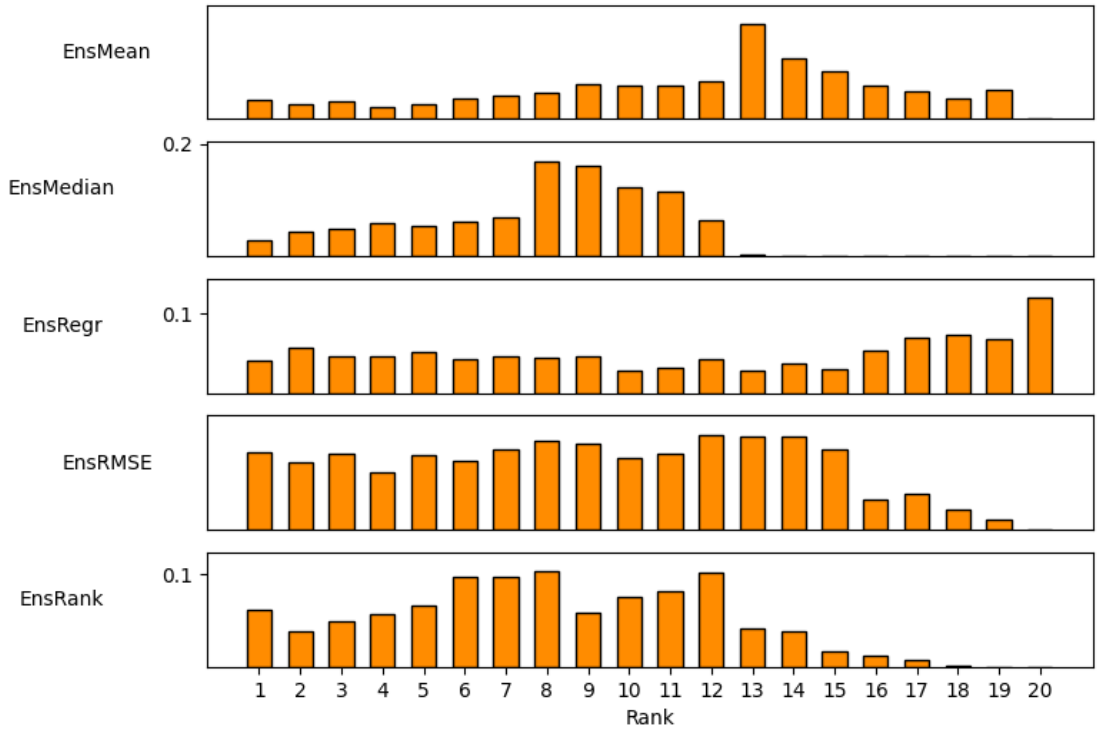



**Figure 4.8:** Histograms of ensemble ranks, ranked by RMSE, for 14-day forecasting on the CovaSim dataset

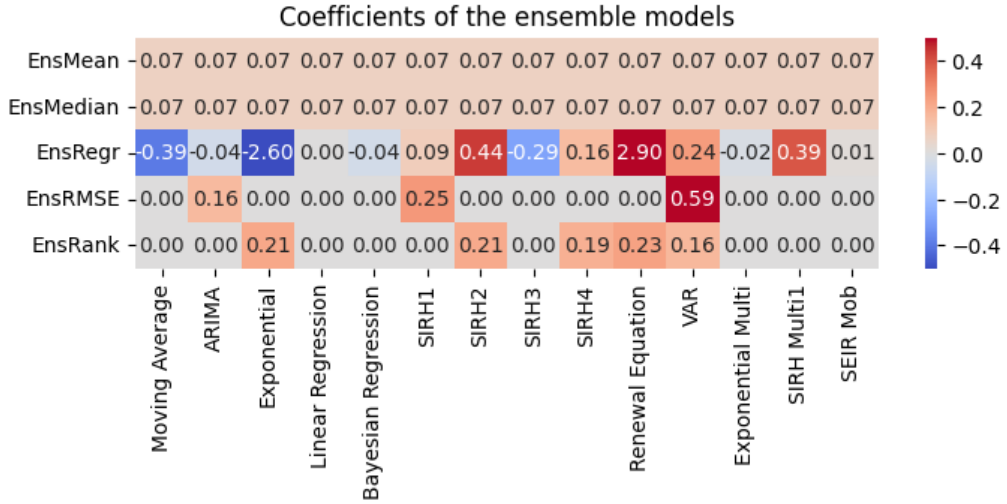



**Figure 4.9:** Coefficients of the ensemble models evaluated on the Norwegian counterfactuals dataset. Model coefficients are fitted on half of the forecasts and ensemble performance is evaluated on the unseen other half.

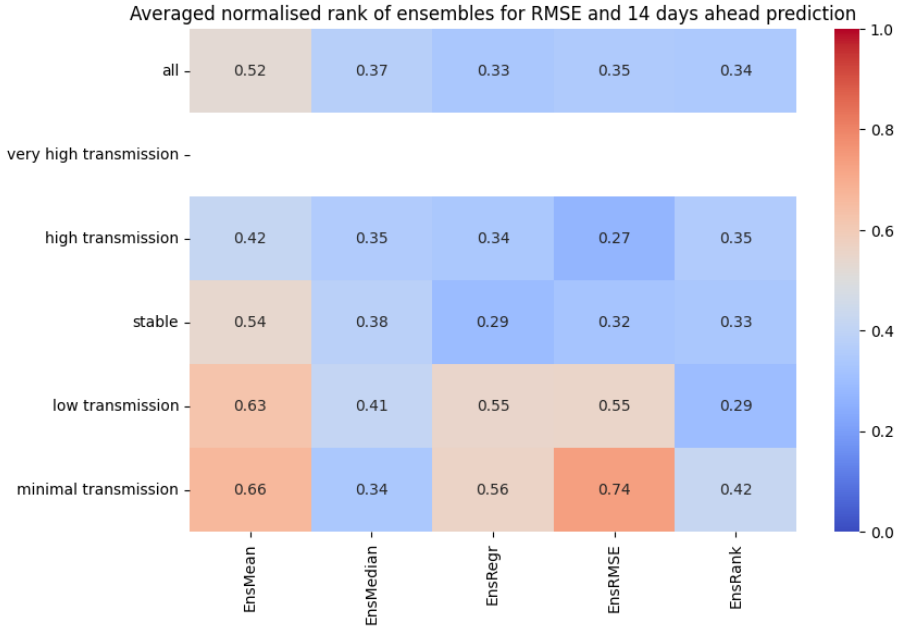



**Figure 4.10:** Average normalized rank of the ensembles across transmission regimes on the Norwegian counterfactuals dataset. Note that no instances of very high transmission occurred in the time series.

### 4.2.2 Norwegian counterfactuals

Even though the relative performance of the component models remained similar between the datasets, the weighted ensembles and their performance differ substantially on the Norwegian counterfactuals. Comparing the ensemble weights in figures 4.5 and 4.9 the allocated weights differ substantially and models included on the CovaSim data are not included when trained on the Norwegian counterfactual time-series. Figure 4.10 shows a similar performance for EnsMedian, EnsRegr, EnsRMSE, and EnsRank where EnsRegr obtains the lowest rank.

Similarly to the CovaSim dataset the Renewal Equation predictor produces forecasts on par with the median (and the other low ranking ensembles) in terms of normalized rank but on this dataset the median ensemble has no clear advantage in the tail, as the tail distributions of both predictors are similar as evident by figure 4.11.

## 4.3 Biases and model correlation

As described in the signal model we partition the model error around the uncertainty of the outcome $Y$ into a bias term $b_i$ and an error term $\varepsilon_i$. In this section we present the estimated biases and study the error distribution as described in section 3.5.

Figure 4.12 presents the bias terms for each model on the CovaSim dataset. Across all forecasts the biases for most model are small with regard to the range of hospitalizations in the data. In general, there is a trend towards underestimation during high transmission and overestimation during low transmission and the SIRH Multi2 model stands out due to its large bias in the low transmission regimes. The biases are similar on the Norwegian counterfactuals, see appendix B.

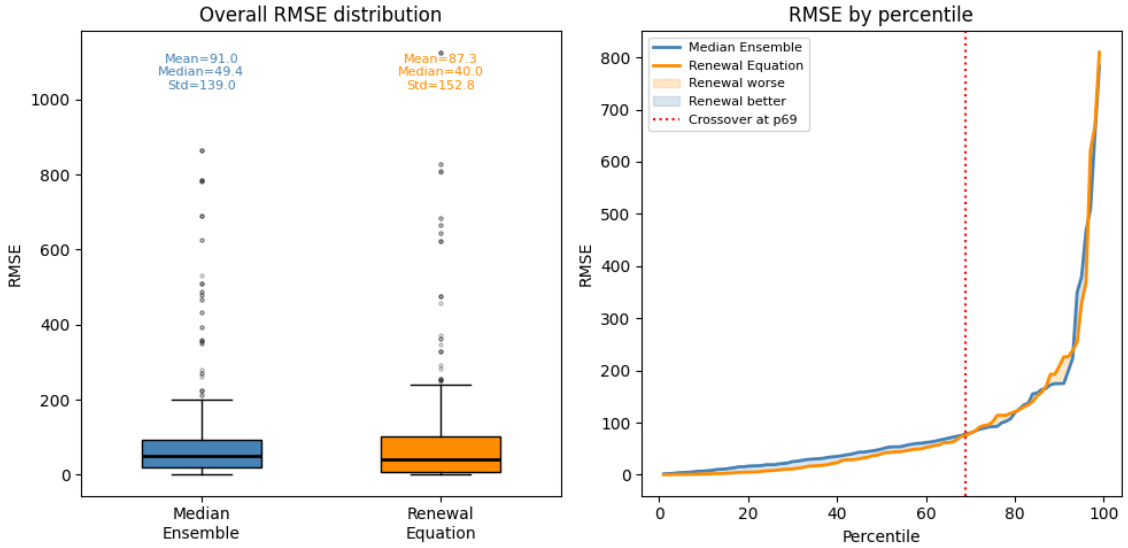



**Figure 4.11:** Comparison of the tail distributions of RMSE for the Renewal Equations and the median ensemble on the Norwegian counterfactual dataset.

**Table 4.2:** Estimated biases of the median and mean ensembles on the CovaSim dataset

| | CovaSim | | Norwegian | |
|---|---|---|---|---|
| | Median bias | Mean bias | Median bias | Mean bias |
| minimal transmission | 634.98 | 967.43 | 7.05 | 25.08 |
| low transmission | 140.96 | 350.34 | 14.39 | 23.26 |
| stable | -1.50 | 21.04 | 30.30 | 25.96 |
| high transmission | -137.49 | -131.15 | -49.32 | -46.24 |
| very high transmission | -34.67 | -86.82 | — | — |
| Overall | 2.88 | 36.53 | 12.01 | 10.59 |

The biases for the Median and Mean ensembles are shown in table 4.2 and the same trend is found here on both datasets, which is to be expected since the ensembles are created from the component models. Overall, both ensembles can be regarded as approximately unbiased which indicate that the error variance is the main driver of ensemble error.

The residual distributions on the CovaSim data are shown in Figure 4.13, from which it is evident that all component model residuals are broadly symmetric with heavier tails than the normal distribution, centred around their respective biases. This is consistent with the contamination model presented in Section 3.4, where a minority of severely erroneous forecasts inflate the tails of an otherwise well-behaved error distribution.

Figure 4.14 shows the pairwise correlations between component model residuals, from which several clear patterns emerge. Most notable is the high correlation among the four SIRH model variants, with similar patterns, albeit less extreme, appearing between other models. The correlation structure is also different when stratified over $R_t$, as is evident by figure 4.15. This indicates that there is information contained within the ensemble spread regarding transmission dynamics which could possibly be exploited for inference.

The correlation structure furthermore suggests that including multiple variants of the same

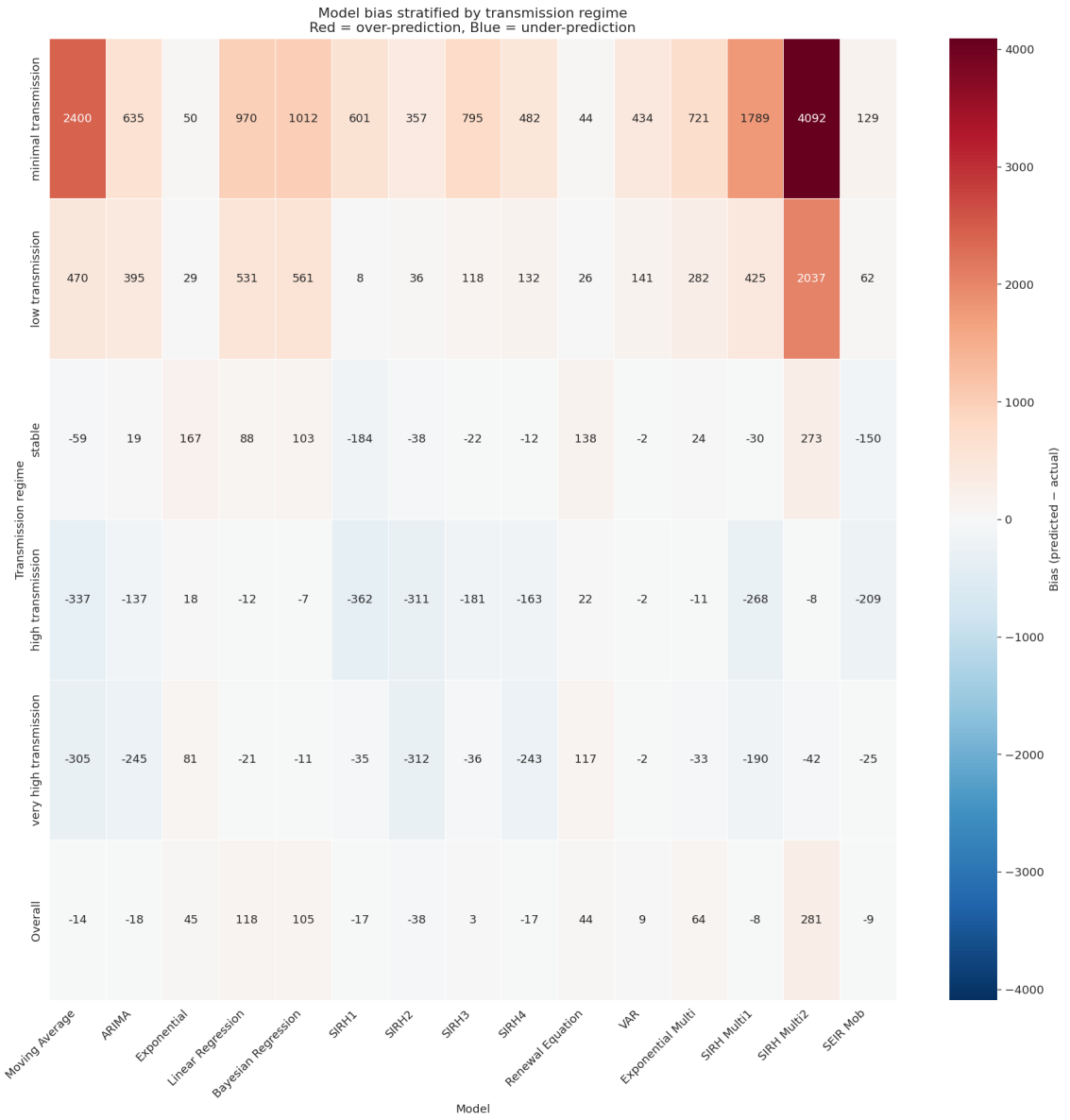



**Figure 4.12:** Estimated model biases by regime on the CovaSim dataset

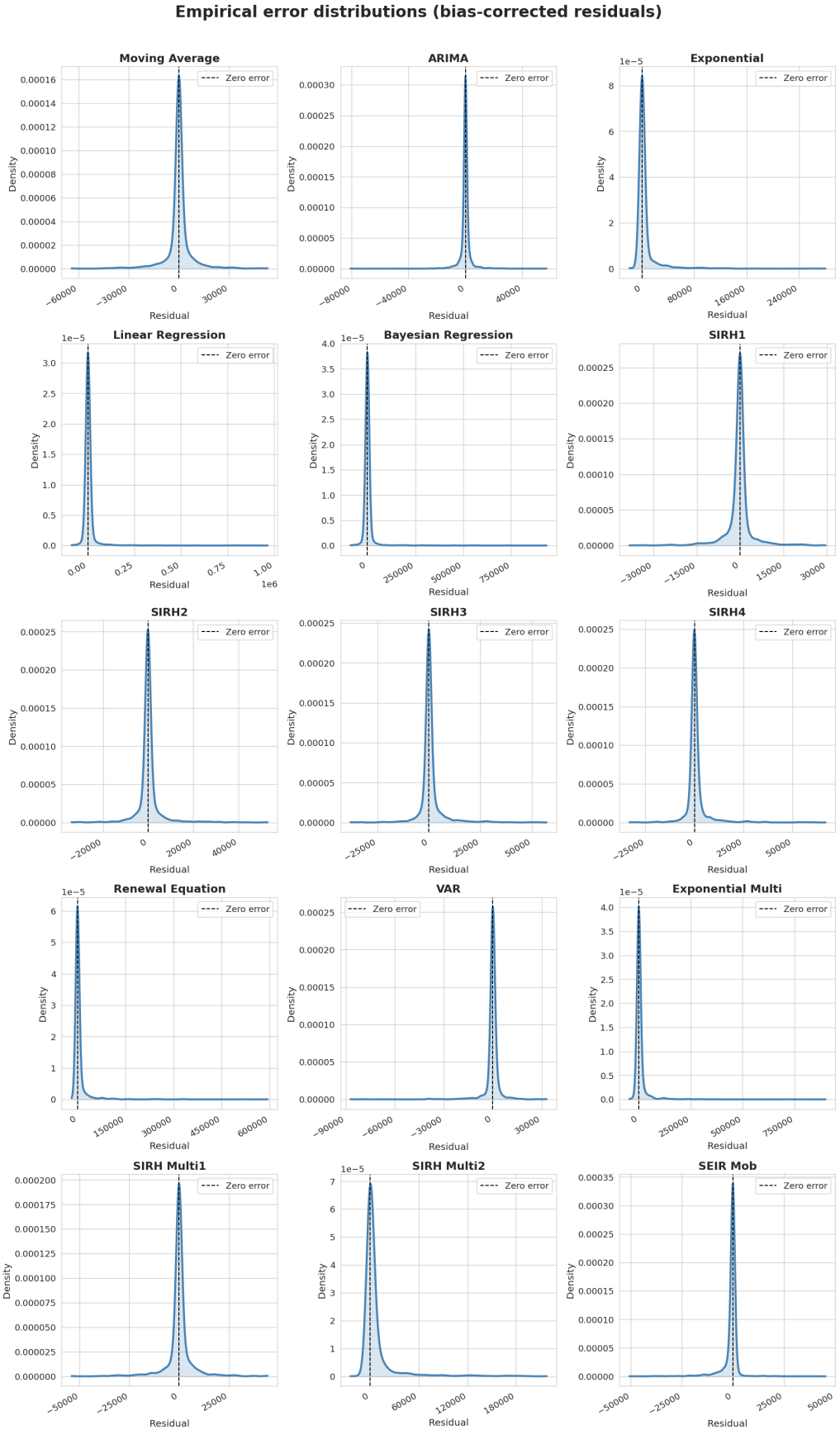



**Figure 4.13:** Residuals of the models on the CovaSim dataset.

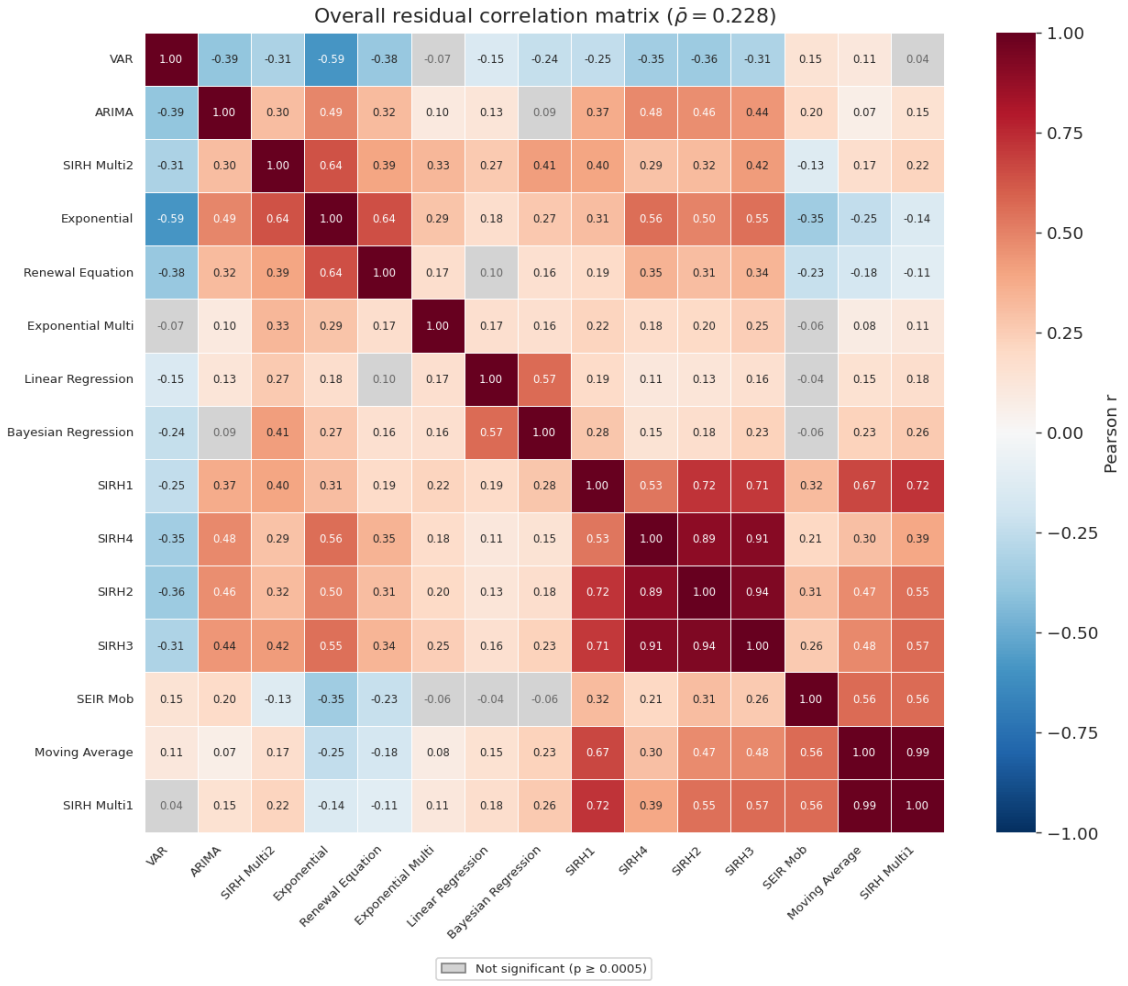



**Figure 4.14:** Residual correlation of the models on the CovaSim dataset

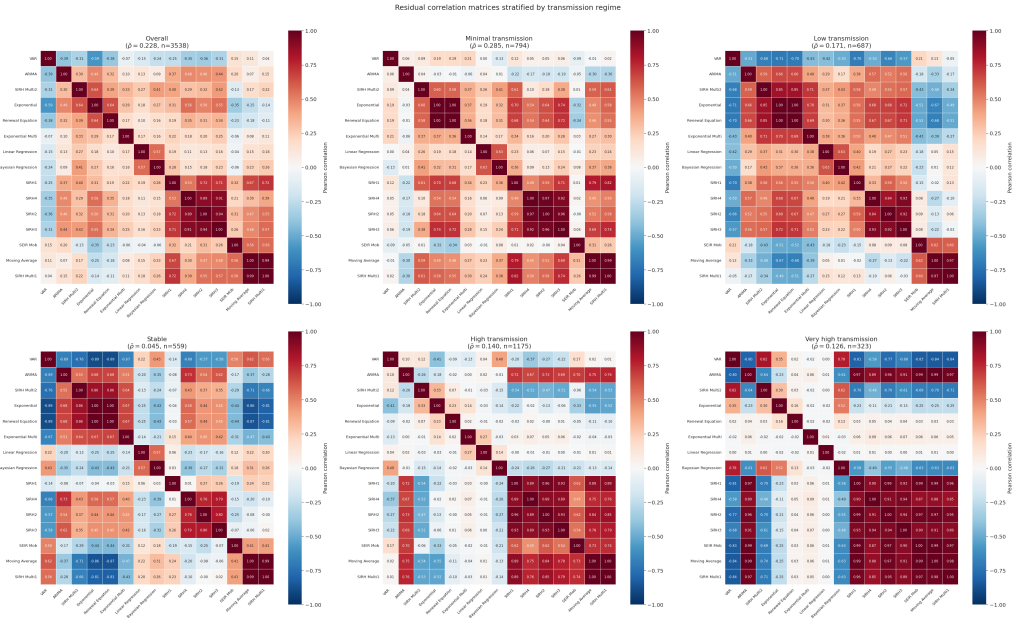



**Figure 4.15:** Residual correlation of the models on the CovaSim dataset stratified over $R_t$

model architecture contributes less to the ensemble, as the shared structural assumptions cause their forecasts to carry largely redundant information. Under the signal model of Section 3.4, this corresponds to a high pairwise correlation $\rho$ among the affected members, which raises the irreducible variance floor and limits the benefit of aggregation as shown in equation 3.18.

This effect is investigated in figure 4.16 where median ensembles of different sizes are created by sub-sampling from the 15 models and their RMSE is plotted against the average correlation within the ensemble on the CovaSim data. The sub-sampled ensembles containing all 4 SIRH models are highlighted in red and a lower envelope is created to show the best available performance for each correlation level. Although the worst ranking ensembles are found across all correlation-levels there is a clear increase in the lowest attainable RMSE as the correlation increases. We interpret this as follows: Model diversity is necessary to produce accurate ensemble forecasts but the ensemble components themself also need to produce good forecasts.

Turning our attention to the Norwegian Counterfactuals, we observe in figure 4.17 that the models overall are much more correlated compared to the CovaSim data, indicating that most ensemble members produce similar errors and hence resulting in a increased average residual correlation.

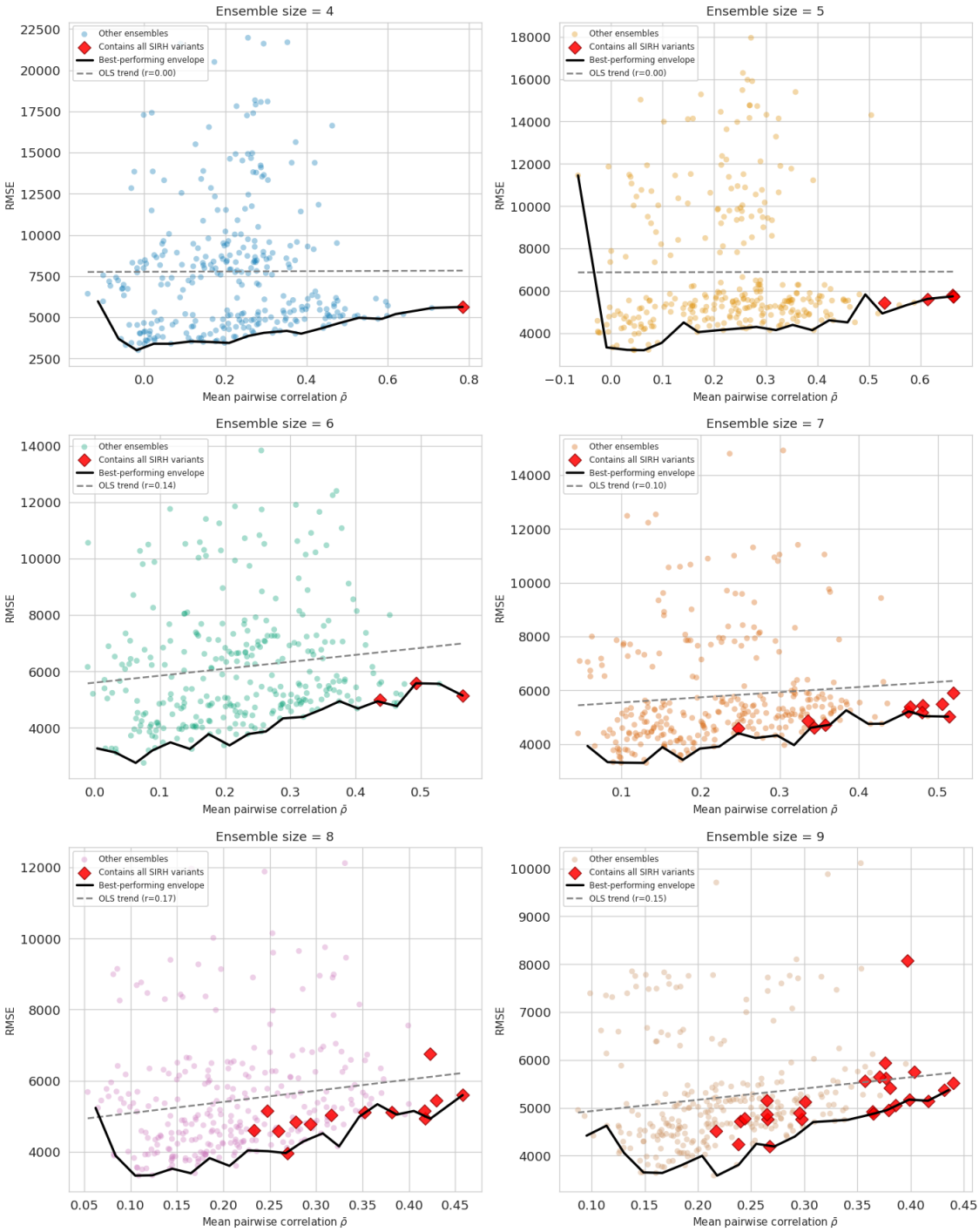



**Figure 4.16:** Ensemble RMSE versus average within-ensemble correlation for sub-sampled median ensembles drawn from the 15 component models on CovaSim data. The lower envelope reveals that while poor ensembles occur at all correlation levels, the best attainable RMSE increases with correlation.

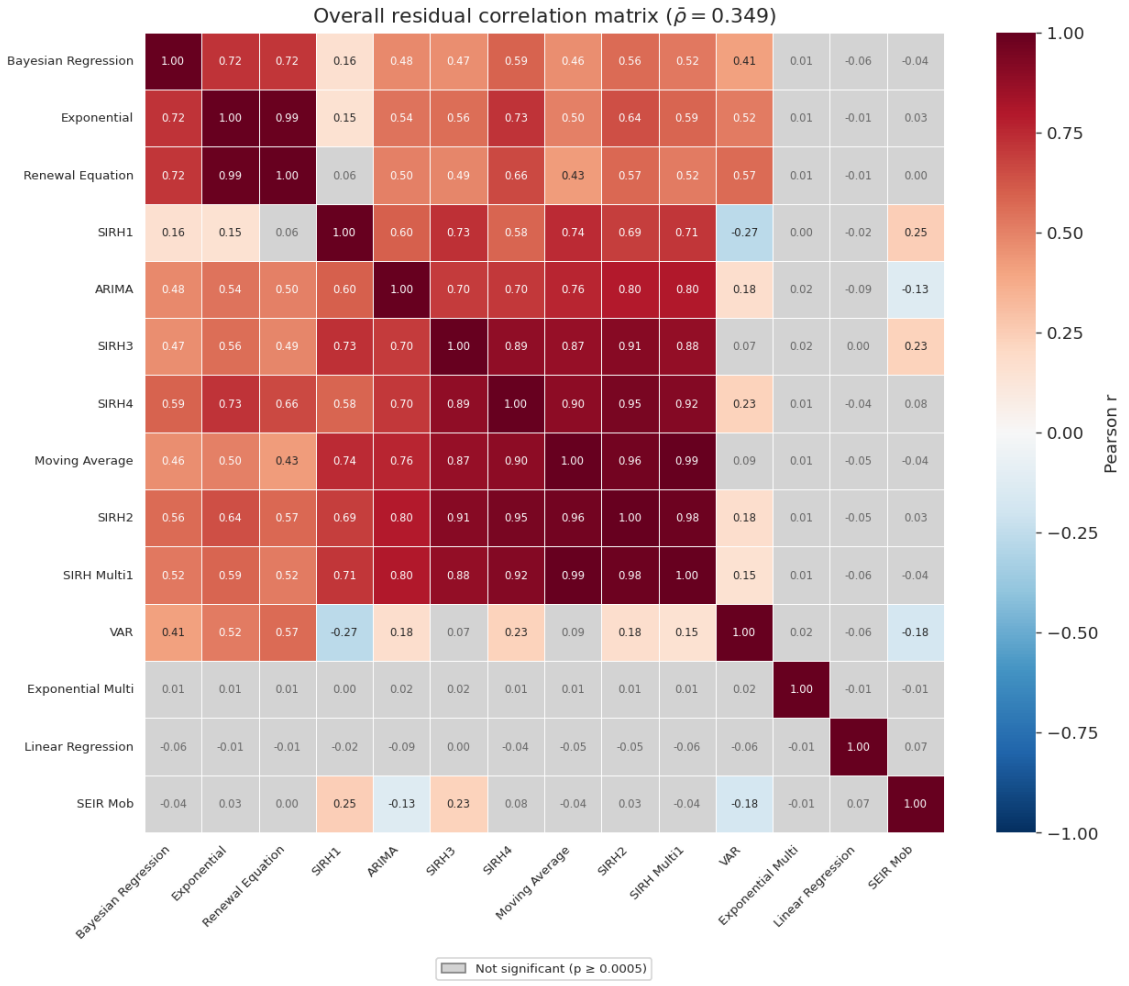



**Figure 4.17:** Residual correlation of the models on the Norwegian counterfactual dataset

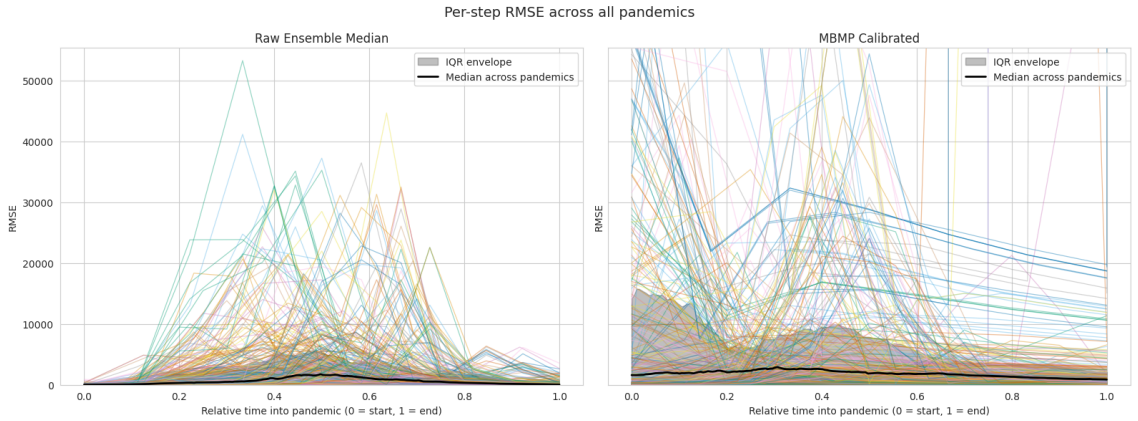



**Figure 4.18:** Per pandemic member-by-member calibration. Each line represents predictions on on pandemic trajectory in the CovaSim dataset

## 4.4 Member-by-member calibration

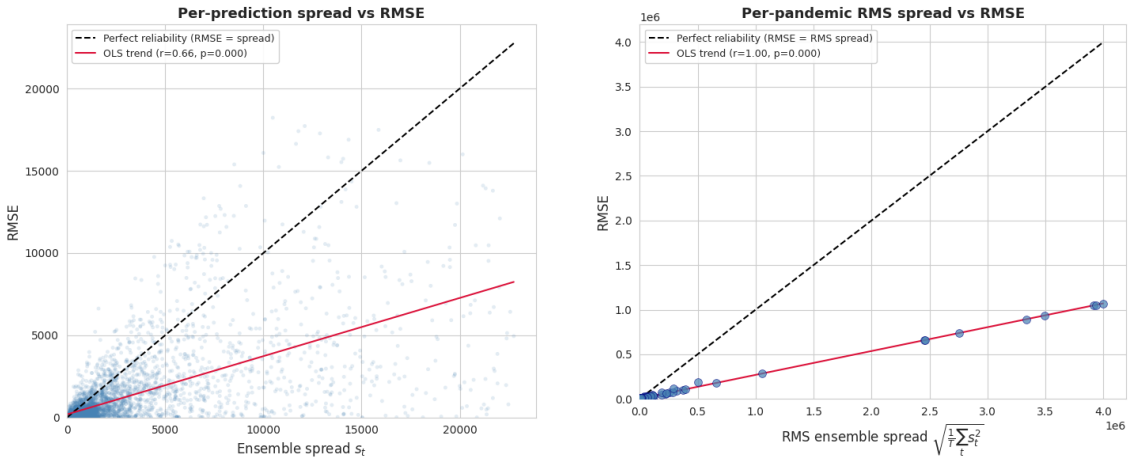
In section 3.6 we introduced Member-by-Member-Post processing which we applied to both dataset in two ways:

- On one pandemic at a time, mimicking a novel infectious disease outbreak.
- Sequentially to all pandemics, mimicking learning from past seasons of a seasonal infectious disease outbreak.

In figure 4.18 the RMSE for each prediction in the CovaSim dataset is showed. The RMSE of the MBMP-ensemble stabilises the further into the pandemic it gets but it only beats the median ensemble 20.4% and 32.2% of the time on the CovaSim and Norwegian counterfactual datasets respectively. On average it produces a reduction in forecast accuracy and is hence not preferable to the median ensemble. Evaluating over several "seasons" does not result in increased performance and the median outperforms MBMP in this case as well.

## 4.5 Spread-skill relationship

Under the reliability assumption presented in Section 2.2.2, a perfectly calibrated ensemble would satisfy RMSE $\approx \sqrt{\frac{1}{T}\sum_{t=1}^{T} s_t^2}$, meaning the RMSE should be proportional to the ensemble spread. A similar relationship emerges from the theoretical framework developed in Section 3.4: under the signal model, the MSE of the median ensemble is given by $\sigma^2 + \frac{\pi\sigma_\epsilon^2}{2N} + \tilde{b}^2$, where the model uncertainty $\sigma_\epsilon^2$ and the median bias $\tilde{b}^2$ act as additive offsets on top of the irreducible outcome uncertainty $\sigma^2$. This implies that the ensemble spread alone overstates the total forecast error, predicting an overdispersed ensemble. The rank histogram analysis in Section 4.1 confirmed that the ensemble is not reliable, which is consistent with this theoretical prediction. We test these results empirically by plotting RMSE against the ensemble spread for each individual prediction and the average spread for each pandemic trajectory in Figure 4.19. The per pandemic RMSE is calculated as the average over median aggregate errors and the per pandemic root-mean-squared (RMS) spread is calculated as $\sqrt{\frac{1}{T}\sum_{t=1}^{T} s_t^2}$ where $s_t$ is the estimated standard deviation of the ensemble at time $t$. The results indicate an overdispersed spread-skill relationship




**(a)** Per forecast **(b)** Per pandemic

**Figure 4.19:** Spread-skill relationship: RMSE plotted against the ensemble spread on the CovaSim data

as predicted, with the ensemble spread consistently exceeding the RMSE, and that the relationship is noisy when evaluated on single forecasts but tightens considerably when averaged across pandemic trajectories.

# 5 | Discussion

This thesis set out to provide a theoretical framework for multi-model ensemble forecasting of infectious disease hospitalisations and to evaluate it empirically across diverse simulated outbreak trajectories. The results broadly support the recommendation of the unweighted median ensemble as a robust default aggregation strategy, consistent with findings from real-world forecasting hubs [4, 13, 14]. However, the picture is not entirely straightforward: while the median ensemble outperforms all weighted alternatives on the CovaSim dataset, the weighted ensembles obtain marginally better average normalised ranks on the Norwegian counterfactual dataset (0.33, 0.34, and 0.35 respectively, compared to 0.37 for the median). Rather than undermining the case for the median, we argue that this difference is informative about the conditions under which trained ensembles can be expected to succeed or fail, and points to a more nuanced understanding of ensemble design than a simple recommendation for any single aggregation method. In the following, we discuss the theoretical and empirical results in turn, examining what they reveal about the role of model diversity, the limits of weight training, and the behaviour of the ensemble across different transmission regimes.

The median ensemble's strong performance on the CovaSim dataset can be understood through the theoretical framework developed in Section 3.4. The key insight is that the median's advantage over the mean is not merely a matter of statistical efficiency, but of robustness: under the contamination model the median's MSE is independent of the noise variance of outlying members as long as fewer than half the ensemble members are outliers, while the mean's MSE grows without bound as the outlying members become more extreme. The empirical residual distributions in Figure 4.13 confirm that this is precisely the setting encountered in practice: all component model residuals exhibit heavier tails than the normal distribution, consistent with a minority of severely erroneous forecasts inflating the tails of an otherwise well-behaved error distribution.

The rank distributions in Figure 4.6 provide further empirical support for this interpretation. The mean ensemble is strongly skewed towards higher ranks, indicating that it occasionally produces very poor forecasts, while the median ensemble's rank distribution is more concentrated. This is the expected signature of contamination: the mean is pulled by the outlying members on the occasions when they produce extreme forecasts, while the median remains anchored to the bulk of the ensemble. The practical consequence is visible in Figure 4.7, where the median ensemble and the renewal equation model achieve similar average normalised ranks on the CovaSim dataset, yet the renewal equation model's RMSE distribution has a substantially heavier upper tail, with the two models performing comparably up to the 73rd percentile before the renewal equation's errors grow substantially larger.

It is worth acknowledging the theoretical cost of this robustness. As shown in Section 3.4, the median incurs a $\pi/2 \approx 1.57$ efficiency penalty relative to the mean under Gaussian noise, meaning that in a setting where all component models produce well-behaved errors, the mean would be the superior aggregator. However, the empirical evidence presented here suggests that such a setting does not arise in practice for infectious disease hospitalization forecasting: the heavy-tailed residual distributions make the robustness of the median more valuable than the efficiency of the mean, and the practical gains from avoiding catastrophic failures outweigh the theoretical efficiency loss. It should be noted, however, that the decomposition of ensemble error into outcome uncertainty $\sigma^2$ and model noise $\sigma_\varepsilon^2$ assumed in the signal model cannot be operationalised from the prediction errors alone. Since each residual $X_{i,t} - Y_t$ contains both the stochastic component of the outcome and the model-specific noise, the two quantities are not separately identifiable from a single observed realisation of $Y_t$. The empirical residual distributions in Figure 4.13 therefore reflect the combined magnitude $\sigma^2 + \sigma_\varepsilon^2$ rather than either term in isolation, and the irreducible floor $\sigma^2$ cannot be estimated without additional assumptions about the data-generating process. This means the theoretical MSE expressions in Table 3.1 should be understood as a qualitative guide to the relative behaviour of aggregators rather than quantities that can be directly verified from the observed residuals.

## 5.1 Weighted vs. unweighted ensembles

The contrast in ensemble performance between the two datasets provides insight into the conditions under which trained ensembles can be expected to succeed or fail. On the CovaSim dataset, the median ensemble outperforms all weighted alternatives, while on the Norwegian counterfactual dataset the weighted ensembles obtain marginally better average normalised ranks. We argue that this difference reflects a fundamental distinction in the structure of the two datasets rather than a general advantage of weight training.

The CovaSim dataset varies the underlying disease biology across 81 parameter combinations, producing a wide range of transmission dynamics with substantially different epidemic shapes. As shown in Figure 4.2, model rankings are highly unstable across these regimes: a model that performs well during high transmission may perform poorly during the declining phase of the same epidemic, and no single model consistently outperforms the others across all regimes. This instability means that weights estimated from past performance are a poor guide to future performance, since the conditions that made a model perform well historically may not persist into the forecast period. Under these conditions the theoretical result of Section 3.4 applies: the optimal population weights are uniform whenever the biases are equal, and unequal weights can only reduce the bias term at the cost of increasing the variance term. When the weights must additionally be estimated from a limited training set, the estimation uncertainty further erodes any potential gains from weighting, consistent with the observation by Ray et al. [13] that trained ensembles struggle when component model skill varies over time.

The Norwegian counterfactual dataset presents a more favourable setting for weight learning. Since the varied parameters relate exclusively to public health interventions (the timing of booster rollout, adherence to self-isolation, and the timing of societal reopening) rather than to the underlying disease biology, the resulting trajectories are more stationary in character. The transmission dynamics of SARS-CoV-2 remain fixed across scenarios, meaning that the relative strengths and weaknesses of the component models are more consistent across trajectories which provides a more stable foundation for cross-trajectory

weight learning, allowing the trained ensembles to identify and exploit systematic differences in component model performance.

This interpretation is further supported by the failure of the member-by-member postprocessing approach on both datasets: since MBMP relies on estimating a stable relationship between ensemble spread and forecast error from past observations, the non-stationary nature of infectious disease time series means that the parameters estimated from early in the epidemic or on previous seasons generalise poorly but the proportion of time the member-by-member postprocessing approach outperformed the median was higher on the norwegian counterfactual dataset, albeit the median ensemble remains preferable in both settings.

Taken together, these results suggest that the choice between trained and untrained ensemble methods should be guided by the stability of the forecasting environment rather than by the size of the available training set alone. In novel outbreak settings where the transmission dynamics are uncertain and evolving, the unweighted median remains the safer choice. In more controlled settings where the epidemic structure is well understood and component model performance is stable, trained ensembles may offer modest but consistent improvements.

## 5.2 The necessity of diversity

A central theoretical prediction of the framework developed in Section 3.4.3 is that ensemble performance is fundamentally limited by the pairwise correlation among component model errors: as shown in equations 3.18 and 3.28, both the mean and median ensemble converge to the same irreducible variance floor $\sigma^2 + \rho\sigma_\epsilon^2$ as the ensemble grows. The empirical correlation analysis confirms that this floor is non-trivial in practice, with a mean pairwise correlation of $\bar{\rho} = 0.228$ on the CovaSim dataset and $\bar{\rho} = 0.349$ on the Norwegian counterfactual dataset.

The most striking feature of the correlation structure is the cluster of high pairwise correlations among the four SIRH model variants. Since the variants share the same underlying model architecture and differ only in whether the recovery rates are treated as free parameters, it is plausible that their forecasts carry partially redundant information, potentially limiting their collective contribution to the ensemble relative to their number.

Figure 4.16 reveals a broader pattern that nuances the diversity argument: while the lower RMSE envelope increases with mean pairwise correlation across all ensemble sizes, the relationship between correlation and RMSE is noisy. We interpret this as follows: model diversity is necessary to produce accurate ensemble forecasts, but the ensemble components themselves also need to produce good forecasts individually. Low correlation alone does not guarantee good performance, and the two requirements — diversity and individual skill — are complementary rather than substitutable. This aligns with the wisdom-of-crowds literature, where [15] caution that group deliberation can improve individual judgements while simultaneously worsening the collective output, precisely because coordinating too closely risks amplifying rather than cancelling individual errors at the aggregate level. Forecast hub modellers must therefore produce skilful forecasts whose errors are wrong *differently* to other participants. From a managerial perspective, this suggests that hub coordinators should resist the temptation to consciously, or unconscionably through close collaboration, homogenise modelling approaches across teams as such con-

vergence may reduce ensemble diversity and degrade aggregate forecast performance even when individual model skill is maintained.

## 5.3 Future research

Several directions for future research emerge naturally from the results of this thesis. A recurring finding is that the correlation structure among component models varies substantially across transmission regimes, and that the ensemble spread carries information about forecast uncertainty through the spread-skill relationship, albeit noisily when evaluated in single forecasts. Together these suggest a potential route to improved ensemble performance: rather than applying fixed weights, one could exploit the regime-dependent correlation structure to dynamically adjust the ensemble composition or weights based on the current transmission context. However, it remains unclear how to translate the observed correlation and spread signals into reliable weight estimates given the non-stationarity of infectious disease time series. This remains an open area for further research. A related theoretical question concerns the accuracy-diversity tradeoff: the results in Figure 4.16 suggest that diversity and individual skill are complementary requirements, but the precise nature of this tradeoff; how much individual accuracy one can sacrifice in exchange for reduced correlation, and vice versa. This has not been formally characterised in the context of multi-model infectious disease ensembles. Developing a principled framework for navigating this tradeoff could provide practical guidance for ensemble design beyond the simple recommendation to include diverse models.

A further natural extension is to move beyond point forecast evaluation and develop a principled approach to probabilistic ensemble construction, addressing the difficulties discussed in Section 2.3. This would require either a method for combining quantile forecasts that correctly widens the ensemble prediction interval when models disagree, or a post-hoc recalibration approach that adjusts the ensemble spread to match the observed spread-skill relationship. Finally, while the use of simulated data allows for controlled evaluation across a wide range of epidemic scenarios, it limits the generalisability of the findings. An important validation step would be to evaluate the theoretical framework and ensemble methods on data from real-world forecasting hubs such as the US COVID-19 Forecast Hub or the European COVID-19 Forecast Hub, where additional challenges such as reporting delays, data revisions, and a heterogeneous pool of contributing models would provide a more demanding test of the conclusions drawn here.

# 6 | Conclusion

This thesis set out to provide a theoretical framework for multi-model ensemble forecasting of infectious disease hospitalisations and to evaluate it empirically across diverse simulated outbreak trajectories. The central finding is that the unweighted median ensemble consistently produces accurate and stable forecasts, performing on par with or better than individual component models and more sophisticated aggregation strategies across both datasets. More broadly, the theoretical framework developed in Chapter 3 shows that the median's robustness to outlying and biased component models, combined with its variance reduction properties, make it a theoretically well-motivated default choice in the setting of infectious disease forecasting where model errors are heavy-tailed, component rankings are unstable across transmission regimes, and the number of available training observations is limited. The empirical results confirm each of these theoretical predictions in turn, and together they provide a coherent explanation for why the median ensemble has been consistently recommended as the aggregation method of choice in real-world infectious disease forecasting hubs [4, 13, 14].

A second contribution of this thesis is the theoretical and empirical characterisation of the role of model diversity in ensemble performance. The framework developed in Section 3.4.3 shows that ensemble error is fundamentally bounded below by the pairwise correlation among component model errors. Notably, the four SIRH model variants exhibit a cluster of particularly high pairwise correlations, suggesting that models sharing the same underlying architecture carry partially redundant information. The sub-sampling analysis further reveals that diversity and individual model skill are complementary rather than substitutable requirements: low correlation alone does not guarantee good ensemble performance, and the practical implication for forecast hub coordinators is that homogenising modelling approaches across teams, whether deliberately or through close collaboration, risks degrading aggregate forecast quality even when individual model skill is maintained.

Several directions for future research emerge naturally from these results. First, the regime-dependent correlation structure observed across transmission contexts, combined with the spread-skill relationship documented empirically, suggests a potential route to improved performance through dynamic ensemble composition or weighting adapted to the current transmission regime. Second, the accuracy-diversity tradeoff identified in the sub-sampling analysis has not been formally characterised in the context of infectious disease ensembles, and a principled framework for navigating this tradeoff would provide practical guidance for ensemble design beyond the simple recommendation to include diverse models.

# Bibliography


[1] F. Ahmed et al., "Systematic review of empiric studies on lockdowns, workplace closures, and other non-pharmaceutical interventions in non-healthcare workplaces during the initial year of the covid-19 pandemic: Benefits and selected unintended consequences," *BMC Public Health*, vol. 24, no. 1, p. 884, 2024. DOI: 10.1186/s12889-024-18377-1.

[2] SVT Nyheter. "Fem år sedan Covid-19: "Borde ha kallat det katastrof från dag ett"," Sveriges Television (SVT). [Online]. Available: https://www.svt.se/nyheter/inrikes/fem-ar-sedan-covid-19-borde-ha-kallat-det-katastrof-fran-dag-ett.

[3] E. Kuhl, *Computational Epidemiology: Data-Driven Modeling of COVID-19* (Springer). Cham, Switzerland: Springer Nature Switzerland AG, 2021, ISBN: 978-3-030-82889-9. DOI: 10.1007/978-3-030-82890-5.

[4] E. Y. Cramer et al., "Evaluation of individual and ensemble probabilistic forecasts of covid-19 mortality in the united states," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 119, no. 15, e2113561119, 2022. DOI: 10.1073/pnas.2113561119.

[5] G. Béchade, T. Lundh, and P. Gerlee, *Evaluation of respiratory disease hospitalisation forecasts using synthetic outbreak data*, 2025. arXiv: 2503.22494 [q-bio.PE]. [Online]. Available: https://arxiv.org/abs/2503.22494.

[6] A. Rodríguez et al., "Machine learning for data-centric epidemic forecasting," *Nature Machine Intelligence*, vol. 6, pp. 1122–1131, 2024. DOI: 10.1038/s42256-024-00895-7.

[7] P. Gerlee et al., "Evaluation and communication of pandemic scenarios," *The Lancet Digital Health*, vol. 6, no. 8, e543–e544, 2024. DOI: 10.1016/S2589-7500(24)00144-4.

[8] H. Thorén and P. Gerlee, "Model uncertainty, the covid-19 pandemic, and the science-policy interface," *Royal Society Open Science*, vol. 11, no. 2, p. 230 803, Feb. 2024, ISSN: 2054-5703. DOI: 10.1098/rsos.230803.

[9] E. Y. Cramer et al., "The united states covid-19 forecast hub dataset," *Scientific Data*, vol. 9, no. 1, p. 462, 2022. DOI: 10.1038/s41597-022-01517-w.

[10] N. G. Reich et al., "Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the u.s.," *PLoS Computational Biology*, vol. 15, no. 11, e1007486, 2019, Published Nov 22 2019. DOI: 10.1371/journal.pcbi.1007486.

[11] C. Viboud et al., "The rapidd ebola forecasting challenge: Synthesis and lessons learnt," *Epidemics*, vol. 22, pp. 13–21, 2018. DOI: 10.1016/j.epidem.2017.08.002.

[12] M. A. Johansson, N. G. Reich, A. Hota, J. S. Brownstein, M. Santillana, et al., "An open challenge to advance probabilistic forecasting for dengue epidemics," *Proceed-*

*ings of the National Academy of Sciences of the United States of America*, vol. 116, no. 48, pp. 24 268–24 274, 2019. DOI: 10.1073/pnas.1909865116.

[13] E. L. Ray et al., "Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States," *International Journal of Forecasting*, vol. 39, no. 3, pp. 1366–1383, 2023. DOI: 10.1016/j.ijforecast.2022.06.005.

[14] F. Becker, K. Sherratt, N. Bosse, and S. Funk, "The influence of ensemble size and composition on the performance of combined real-time COVID-19 forecasts," *medRxiv*, 2025, Preprint. DOI: 10.1101/2025.08.09.25331484.

[15] A. Lyon and E. Pacuit, "The wisdom of crowds: Methods of human judgement aggregation," in *Handbook of Human Computation*, P. Michelucci, Ed., New York: Springer, 2013, pp. 599–614. DOI: 10.1007/978-1-4614-8806-4_47.

[16] H. Rauhut and J. Lorenz, "The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions," *Journal of Mathematical Psychology*, vol. 55, no. 3, pp. 191–197, 2011. DOI: 10.1016/j.jmp.2010.10.002.

[17] R. M. Levenson, E. A. Krupinski, V. M. Navarro, and E. A. Wasserman, "Pigeons (columba livia) as trainable observers of pathology and radiology breast cancer images," *PLOS ONE*, vol. 10, no. 11, e0141357, 2015. DOI: 10.1371/journal.pone.0141357.

[18] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Series in Statistics), 2nd ed. New York: Springer, 2009, ISBN: 978-0-387-84857-0.

[19] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324.

[20] E. N. Lorenz, "Deterministic nonperiodic flow," *Journal of the Atmospheric Sciences*, vol. 20, no. 2, pp. 130–141, 1963. DOI: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

[21] E. N. Lorenz, "A study of the predictability of a 28-variable atmospheric model," *Tellus*, vol. 17, no. 3, pp. 321–333, 1965. DOI: 10.3402/tellusa.v17i3.9076.

[22] J. Du et al., "Ensemble methods for meteorological predictions," in *Handbook of Hydrometeorological Ensemble Forecasting*, Q. Duan et al., Eds., Springer-Verlag, 2018. DOI: 10.1007/978-3-642-40457-3_13-1.

[23] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 2, pp. 243–268, 2007. DOI: 10.1111/j.1467-9868.2007.00587.x.

[24] A. Dirkson and M. Buehner, "Are we misdiagnosing ensemble forecast reliability? on the insufficiency of spread–error and rank-based reliability metrics," *arXiv preprint arXiv:2512.02160*, 2025. [Online]. Available: https://arxiv.org/abs/2512.02160.

[25] J. Bröcker and H. Kantz, "The concept of exchangeability in ensemble forecasting," *Nonlinear Processes in Geophysics*, vol. 18, no. 1, pp. 1–5, 2011. DOI: 10.5194/npg-18-1-2011.

[26] V. Fortin, M. Abaza, F. Anctil, and R. Turcotte, "Why should ensemble spread match the rmse of the ensemble mean?" *Journal of Hydrometeorology*, vol. 15, no. 4, pp. 1708–1713, 2014. DOI: 10.1175/JHM-D-14-0008.1, Corrigendum published in *Journal of Hydrometeorology*, 16 (2015), p. 484, doi:10.1175/JHM-D-14-0161.1.

[27] D. S. Wilks, "On the reliability of the rank histogram," *Monthly Weather Review*, vol. 139, no. 1, pp. 311–316, 2011. DOI: 10.1175/2010MWR3446.1.

[28] J. Bröcker and H. Kantz, "The concept of exchangeability in ensemble forecasting," *Nonlinear Processes in Geophysics*, vol. 18, no. 1, pp. 1–5, 2011. DOI: 10.5194/npg-18-1-2011.

[29] A. Diz-Lois Palomares, B. Freiesleben de Blasio, L. Y. H. Chan, F. Di Ruscio, and J. E. Midtbø, "Combining an agent-based model with Gaussian process emulation to model the emergence of the SARS-CoV-2 Omicron variant in Norway," *Epidemics*, vol. 54, p. 100 895, 2026. DOI: 10.1016/j.epidem.2026.100895.

[30] P. Gerlee et al., "Predicting regional COVID-19 hospital admissions in Sweden using mobility data," *Scientific Reports*, vol. 11, no. 1, p. 24 171, Dec. 2021. DOI: 10.1038/s41598-021-03499-y.

[31] Sam Abbott, Katherine Sherratt, Sebastian Funk, *NFIDD: Nowcasting and Forecasting Infectious Disease Dynamics*, 2025. [Online]. Available: https://nfidd.github.io/nfidd/.

[32] A. Cori, N. M. Ferguson, C. Fraser, and S. Cauchemez, "A new framework and software to estimate time-varying reproduction numbers during epidemics," *American Journal of Epidemiology*, vol. 178, no. 9, pp. 1505–1512, Nov. 2013, ISSN: 0002-9262. DOI: 10.1093/aje/kwt133.

[33] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. New York: John Wiley & Sons, 1987, ISBN: 9780471852331.

[34] Wikipedia, *Pearson correlation coefficient*, Accessed: 2026-05-21, Wikipedia, The Free Encyclopedia, 2025. [Online]. Available: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#Inference.

[35] R. Schefzik, "Ensemble calibration with preserved correlations: Unifying and comparing ensemble copula coupling and member-by-member postprocessing," *Quarterly Journal of the Royal Meteorological Society*, vol. 143, no. 703, pp. 999–1008, 2017. DOI: 10.1002/qj.2984.

# A | Supporting results

**Lemma 1.** *Let* $\mathbf{w} = (w_1, \ldots, w_N)$ *satisfy the linear constraint*

$$\sum_{i=1}^{N} w_i = 1.$$

*Then the Euclidean norm* $\|\mathbf{w}\|$ *is minimized when*

$$w_i = \frac{1}{N}, \qquad i = 1, \ldots, N.$$

*Proof.* Consider the vectors $\mathbf{w} = (w_1, \ldots, w_N)$ and $\mathbf{1} = (1, \ldots, 1)$. By the Cauchy–Schwarz inequality,

$$(\mathbf{w}^\top \mathbf{1})^2 \;\leq\; \|\mathbf{w}\|^2 \, \|\mathbf{1}\|^2.$$

Using the constraint $\mathbf{w}^\top \mathbf{1} = \sum_{i=1}^{N} w_i = 1$ and $\|\mathbf{1}\|^2 = N$, we obtain

$$1 \;\leq\; \|\mathbf{w}\|^2 \, N \qquad \Longrightarrow \qquad \|\mathbf{w}\|^2 \;\geq\; \frac{1}{N}.$$

Equality in Cauchy–Schwarz holds if and only if $\mathbf{w}$ and $\mathbf{1}$ are linearly dependant, i.e. $w_1 = \cdots = w_N$. Imposing the constraint $\sum_{i=1}^{N} w_i = 1$ then yields

$$w_i = \frac{1}{N}, \qquad i = 1, \ldots, N.$$

Thus the equal–weight vector uniquely minimizes the Euclidean norm. □

# B | Results of the Norwegian counterfactual data-set

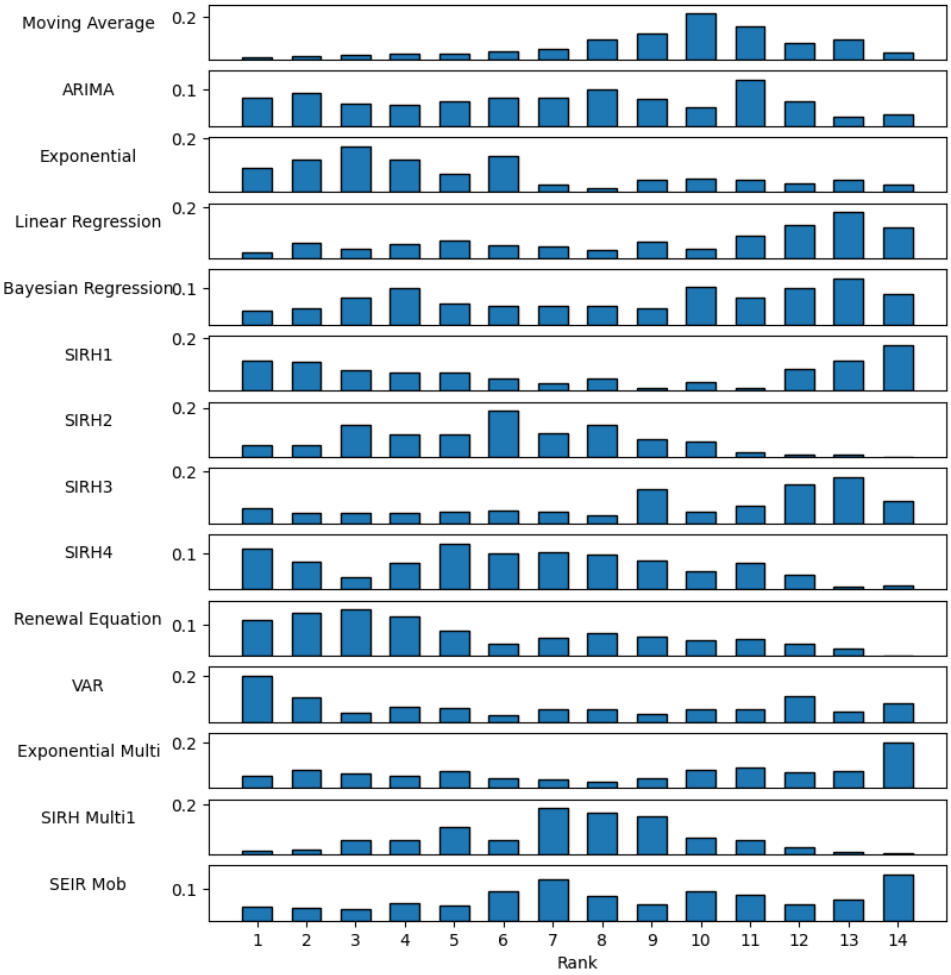



**Figure B.1:** Histograms of model ranks, ranked by RMSE, for 14-day forecasting on the Norwegian counterfactual dataset

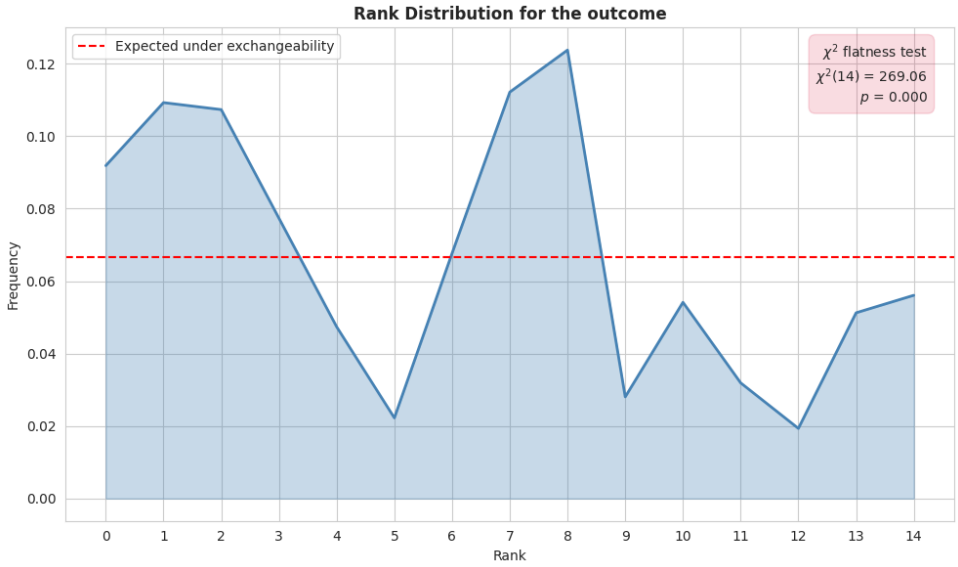



**Figure B.2:** Rank histogram of the outcome on the Norwegian counterfactual dataset.

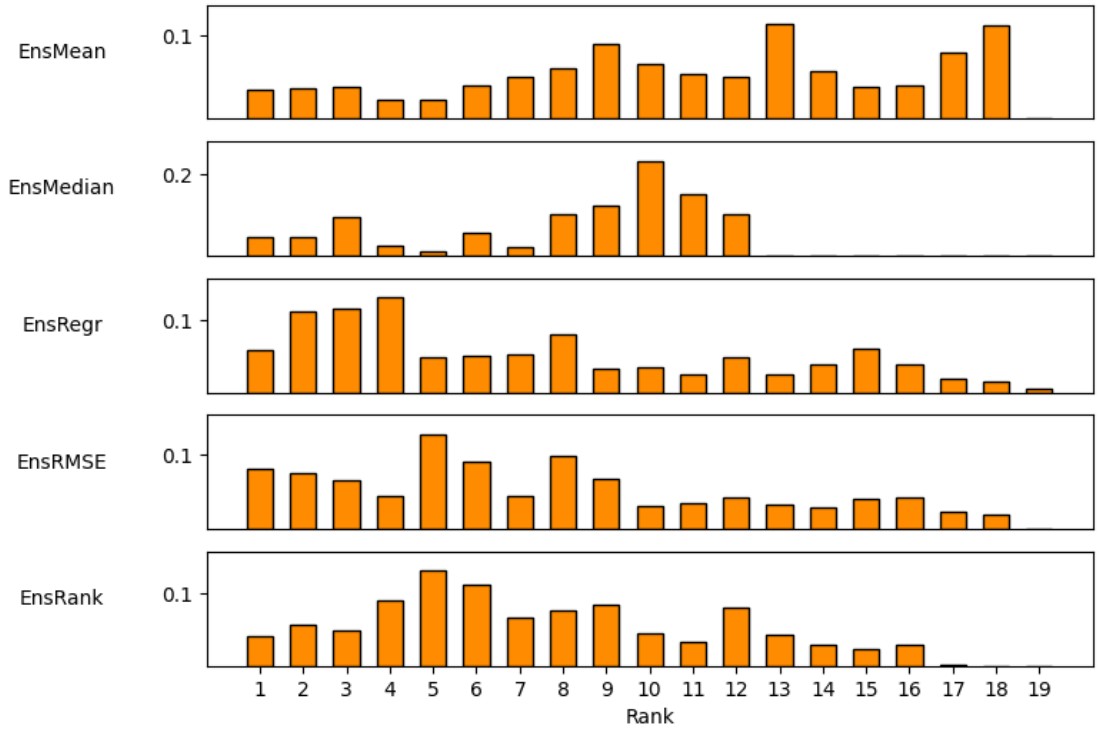



**Figure B.3:** Histograms of ensemble ranks, ranked by RMSE, for 14-day forecasting on the Norwegian counterfactual dataset

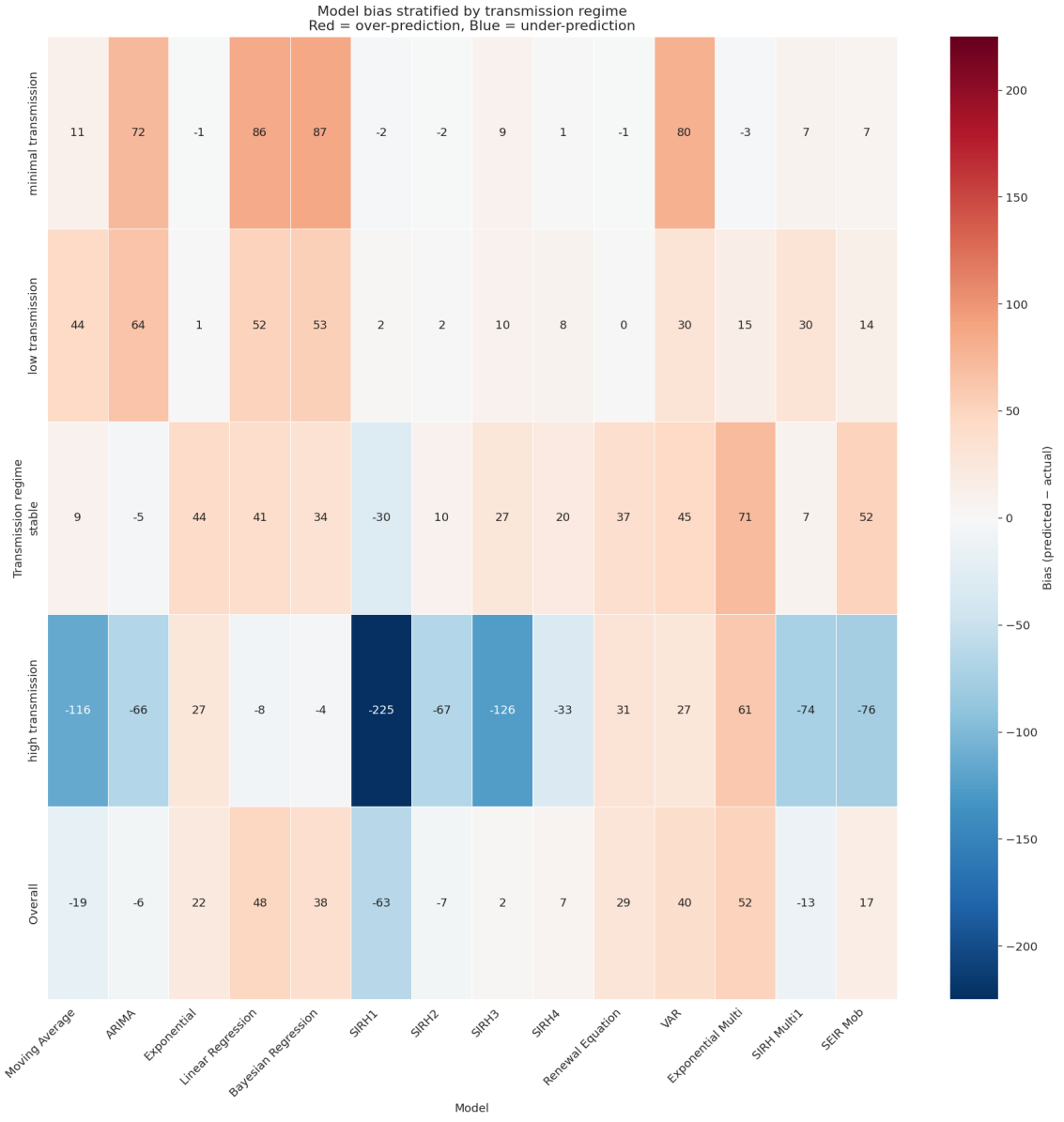



**Figure B.4:** Estimated model biases by regime on the Norwegian counterfactual dataset

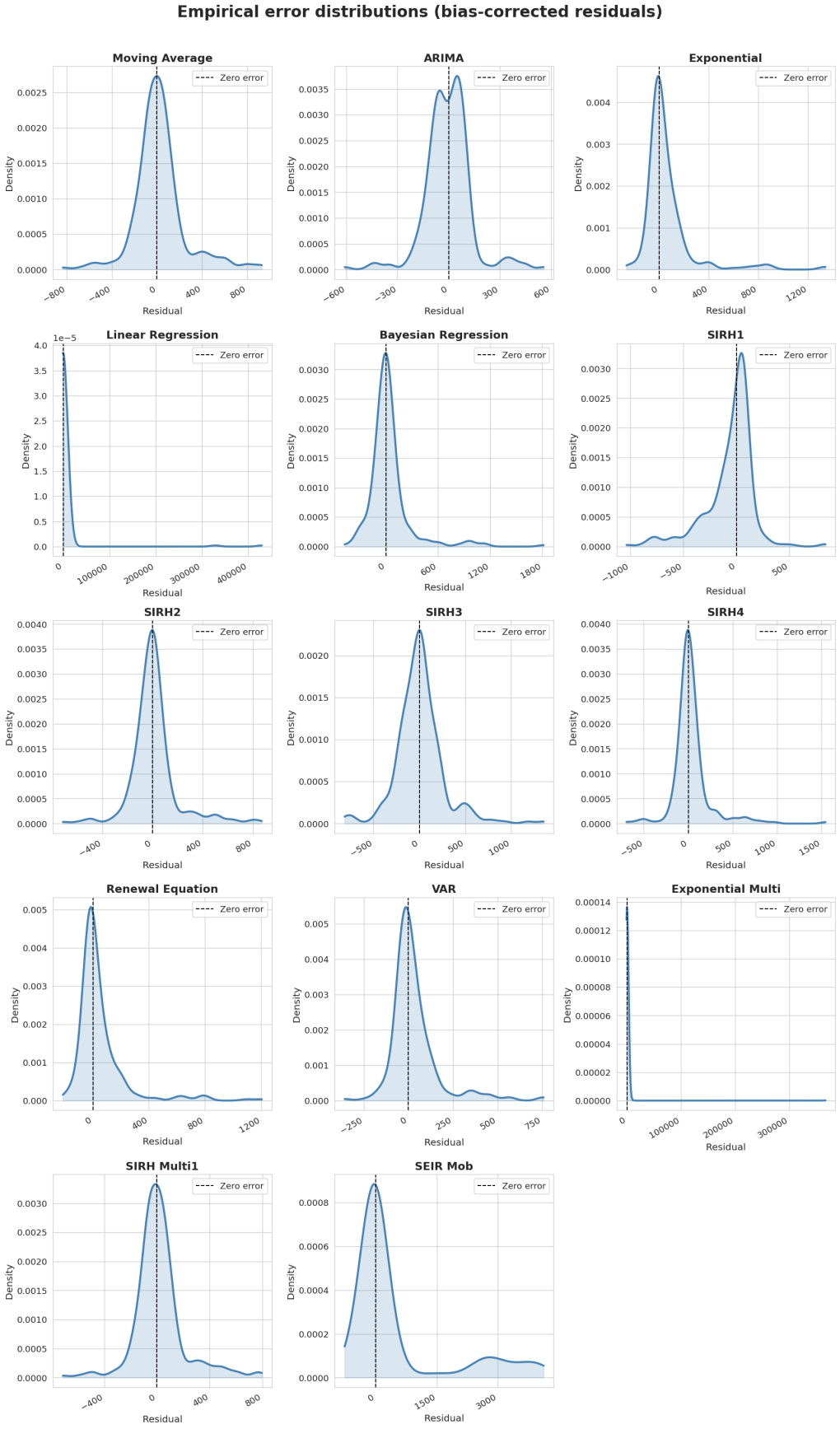



**Figure B.5:** Residuals of the models on the Norwegian counterfactual dataset.

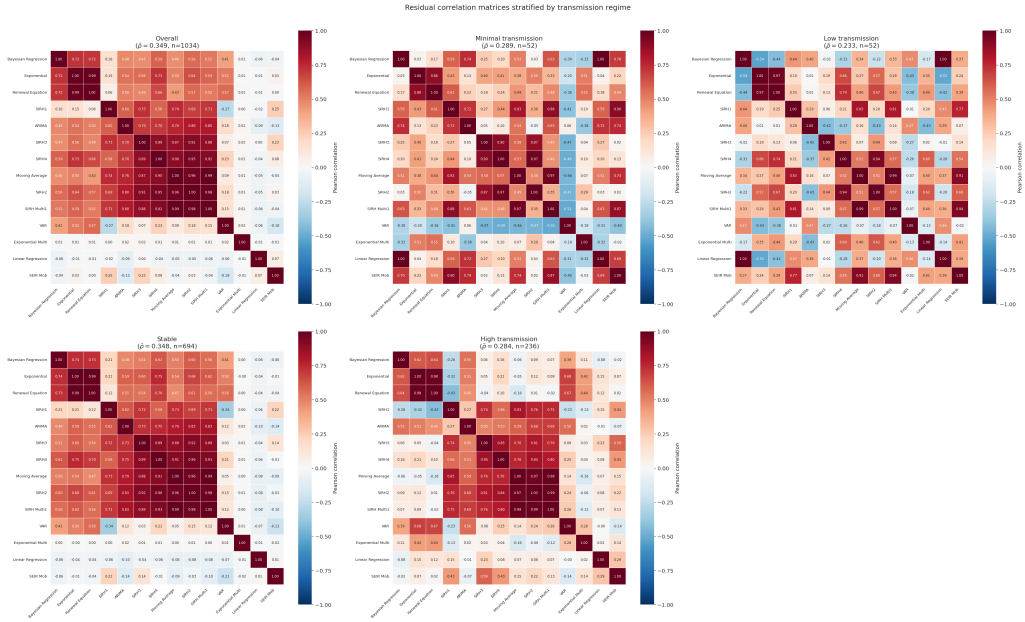



**Figure B.6:** Residual correlation of the models on the Norwegian counterfactual dataset stratified over $R_t$

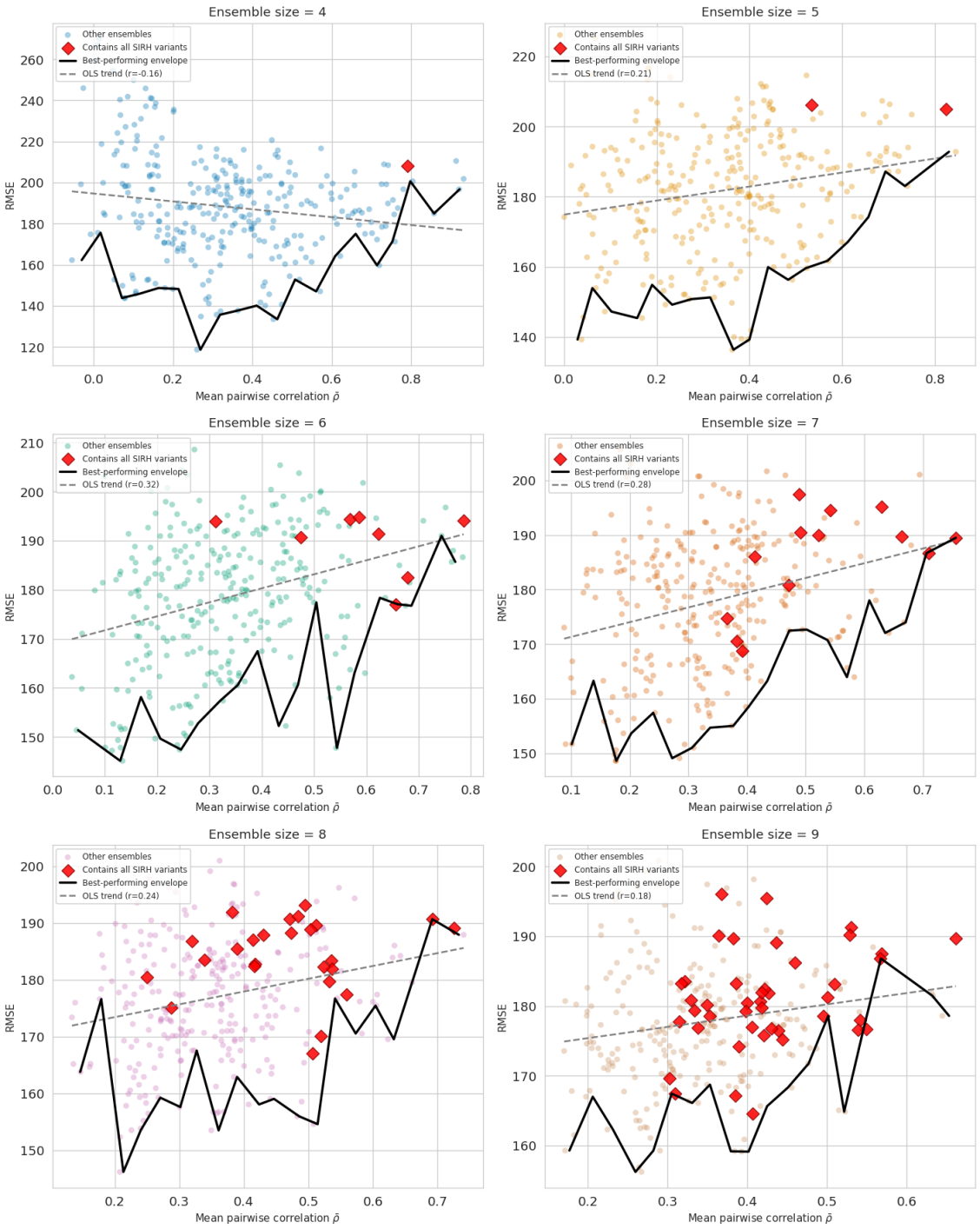



**Figure B.7:** Ensemble RMSE versus average within-ensemble correlation for sub-sampled median ensembles drawn from the 14 component models on Norwegian counterfactual data. The lower envelope reveals that while poor ensembles occur at all correlation levels, the best attainable RMSE increases with correlation.

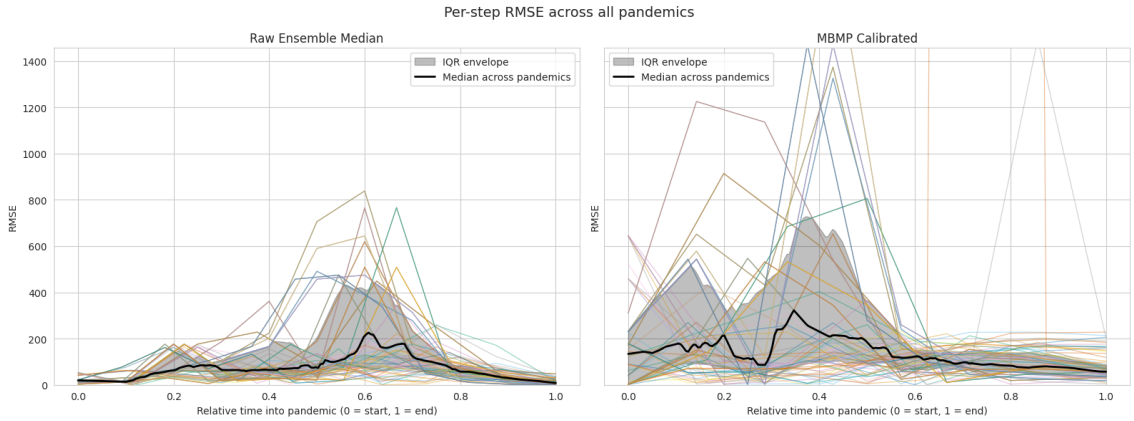
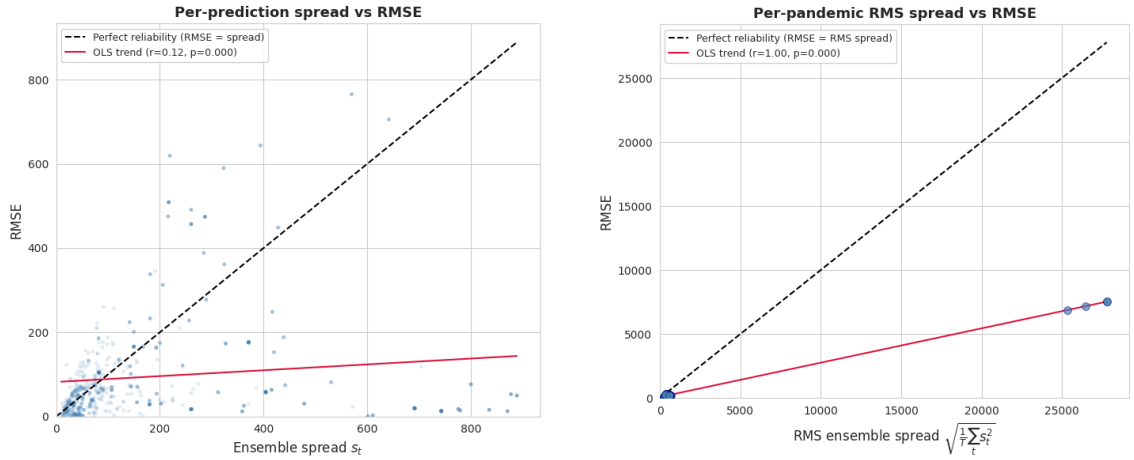



**Figure B.8:** Per pandemic member-by-member calibration. Each line represents predictions on on pandemic trajectory in the Norwegian counterfactual dataset




**(a)** Per forecast **(b)** Per pandemic

**Figure B.9:** Spread-skill relationship: RMSE plotted against the ensemble spread on the Norwegian counterfactual data

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


CHALMERS
UNIVERSITY OF TECHNOLOGY