

Sequential model confidence sets

Sebastian Arnold^{1,†}, Georgios Gavriloopoulos^{2,†} , Benedikt Schulz³
and Johanna Ziegel²

¹Centrum Wiskunde & Informatica (CWI), Science Park 123, 1098 XG Amsterdam, The Netherlands

²Department of Mathematics, Eidgenössische Technische Hochschule Zürich (ETHZ), Seminar for Statistics, Rämistrasse 101, 8092 Zürich, Switzerland

³Karlsruhe Institute of Technology (KIT), Institute of Stochastics, Englerstrasse 2, 76131 Karlsruhe, Germany

Address for correspondence: Georgios Gavriloopoulos, Department of Mathematics, Eidgenössische Technische Hochschule Zürich (ETHZ), Seminar for Statistics, Rämistrasse 101, 8092 Zürich, Switzerland.

Email: georgios.gavriloopoulos@stat.math.ethz.ch

Abstract

In most prediction and estimation situations, scientists consider various statistical models for the same problem, and naturally want to select amongst the best. Hansen et al. [(2011). The model confidence set. *Econometrica: Journal of the Econometric Society*, 79(2), 453–497] provide a powerful solution to this problem by the so-called model confidence set, a subset of the original set of available models that contains the best models with a given level of confidence. Importantly, model confidence sets respect the underlying selection uncertainty by being flexible in size. However, they presuppose a fixed sample size which stands in contrast to the fact that model selection and forecast evaluation are inherently sequential tasks where we successively collect new data and where the decision to continue or conclude a study may depend on the previous outcomes. In this article, we extend model confidence sets sequentially over time by relying on sequential testing methods through e-processes and confidence sequences. Sequential model confidence sets allow to continuously monitor the models' performances and come with time-uniform, nonasymptotic coverage guarantees.

Keywords: forecast comparison, forecast evaluation, model confidence set, multiple testing, sequential inference

1 Introduction

In science as well as in our daily life, we frequently encounter situations in which multiple statistical models or forecasts are available for the same problem and where we have to decide which model(s) or forecast(s) we want to trust. We might think for example of multiple regression models for inflation with respect to different sets of covariates or of multiple weather services which issue precipitation predictions for the next day. In such situations, we naturally want to select the best model(s) or forecast(s), where the term 'best' is defined in terms of a user-specified criterion, typically given by some real-valued loss function where lower scores correspond to better performance.

The *model confidence set* (MCS) proposed by Hansen et al. (2011) provides a promising solution to this problem. It departs from standard practice where just a single model is selected according to some appropriate loss and nothing is said about the uncertainty associated with this selection. The MCS takes this uncertainty into account by reducing the original set of available models to a smaller set of flexible size that contains the best models with a given level of confidence. Model confidence sets are of great importance in applications without an obvious benchmark, may be easily constructed by the MCS algorithm and are widely applied in the econometrics community; see, e.g. Nan (2009) and Peng et al. (2018) amongst others.

[†] S.A. and G.G. contributed equally.

However, even though the MCS provides an appealing solution to a highly important problem, it comes with limitations. In particular, it is assumed that the sequences of observed loss differences are stationary, an assumption which may be questionable in practice since we may expect that the models evolve over time, and correct for errors and systematic biases by using past information. Apart from relying on strong stationarity assumptions, the MCS requires a sample of some fixed size chosen independently of the data. In other words, we assume that we observe losses of the models over some prespecified evaluation period in order to compute the MCS only once at the end of the period. This procedure highly contrasts our natural urge to assess the models on a regular basis by successively including past observations to the available dataset. Consider for example different weather institutions that predict the accumulated precipitation for the following day on a daily basis. Then, firstly, it is not given that the initially better models remain superior until the end of our study. Secondly, we want to assess the forecasters and monitor their performance sequentially, say, at the end of each day, week, or month, since the forecasters issue their forecasts sequentially as well and may behave nonstationarily. If we gather evidence that one institution is worse than the others but the evidence is not yet statistically significant, then we want to continue to collect evidence for this hypothesis without losing the information of the earlier observations. Crucially, we want to decide at the end of each day, week or month whether to continue or conclude the study, depending on the previous outcomes. Finally, the derivation of the MCS is based on bootstrap approximations and asymptotic, Central-Limit-Theorem-related properties. In addition to high computational costs, coverage guarantees are only asymptotic.

Selecting models and comparing forecasts are inherently sequential tasks, and we should seek methods that allow for a sequential choice of the best models while respecting the underlying selection uncertainty. In this article, we address this need and contribute with sequential model confidence sets. Sequential model confidence sets generalize the core idea of [Hansen et al. \(2011\)](#) sequentially over time and enjoy time-uniform coverage guarantees of the theoretically superior objects. Moreover, they have nonasymptotic validity and do not rely on computationally expensive approximations. Our methods rely on e-processes and time-uniform confidence sequences, which lay the foundation for safe anytime-valid inference, a field which met a surge of new contributions over the last years by various authors; see, e.g. [Shafer \(2021\)](#), [Vovk and Wang \(2021\)](#), [Ramdas et al. \(2023\)](#), and [Grünwald et al. \(2024\)](#) amongst others.

Sequential testing methods relying on minimal distributional assumptions have been used before for sequential forecast evaluation: [Henzi and Ziegel \(2022\)](#) provide e-processes to sequentially test the strong hypothesis that one forecaster is better than another at all time points. In contrast, [Choe and Ramdas \(2023\)](#) develop e-processes for the weaker hypothesis that one forecaster is uniformly better than the other on average. Both studies assess the performance of some forecaster at time $t \in \mathbb{N}$ with respect to the (average) expected loss difference given all information up to the previous time step. This is relevant to tests of equal conditional predictive ability proposed by [Giacomini and White \(2006\)](#). In contrast to the unconditional approach, most prominently in the seminal work of [Diebold and Mariano \(1995\)](#), the conditional approach uses the available information to infer which forecaster is more accurate on a specific date. In this article, we follow the conditional approach and study the strong as well as the weak hypothesis. In contrast to all aforementioned contributions on forecast comparison, our methods allow to simultaneously draw inference on the performance of $m \geq 2$ different forecasters rather than only being able to compare two different ones.

In order to construct sequential model confidence sets, we make use of the e-processes provided by [Howard et al. \(2020\)](#) and [Choe and Ramdas \(2023\)](#) under the assumption of sub-Gaussian or sub-exponential loss differences, which subsumes in particular the case where the loss differences are conditionally bounded, an assumption, which does not only hold for bounded losses but also for prominent choices of scoring rules such as the continuous ranked probability score (CRPS; [Matheson & Winkler, 1976](#)) or the quantile score ([Gneiting, 2011](#)).

The remainder of the article is structured as follows. Section 2 presents the problem and introduces sequential model confidence sets. In Section 3, we review important definitions from sequential testing theory and provide methods to construct sequential model confidence sets with respect to three particular notions of superior objects. A simulation study is conducted in Section 4 before we apply the proposed methods in case studies on Covid-19-related deaths and wind gust predictions in Section 5. The main part of the article closes with a discussion in Section 6. Some

background on forecast evaluation, technical comments, and proofs are available in the [online supplementary material](#).

2 Sequential model confidence sets

We consider a set $\mathcal{M}_0 = \{1, \dots, m\}$ indexing $m \geq 2$ different statistical models or forecasters, simply referred to as *models*, and we let L be a loss function which measures the quality of the models and is negatively oriented, that is, lower values correspond to better performance. We collect data in discrete time and denote the loss of model $i \in \mathcal{M}_0$ at time $t \in \mathbb{N}$ by $L_{i,t}$. We assume that all random quantities are defined on some underlying measurable space $(\Omega, \mathcal{F})$ equipped with some filtration $(\mathcal{F}_t)_{t \in \mathbb{N}}$. We denote the family of all probability measures on $(\Omega, \mathcal{F})$ by $\mathfrak{B}(\Omega)$ and write $\mathbb{Q}$ for a generic element in $\mathfrak{B}(\Omega)$.

Our leading example of the above setting is the following. Let $1, \dots, m$ index different forecasters which all sequentially issue predictive distributions or point forecasts $(f_{1,t})_{t \in \mathbb{N}}, \dots, (f_{m,t})_{t \in \mathbb{N}}$ for some unknown quantity $(Y_t)_{t \in \mathbb{N}}$, adapted to $(\mathcal{F}_t)_{t \in \mathbb{N}}$. For each $i = 1, \dots, m$ and $t \in \mathbb{N}$, the forecast $f_{i,t}$, which refers to the outcome Y_t , is *predictable*, meaning that it is only based on information until time $t - 1$. To assess the forecasters, one should employ *proper scoring rules* or *consistent scoring functions*, see [online supplementary material A](#) for further details. Then, the time series $(L_{i,t})_{t \in \mathbb{N}}$ emerges by applying a proper scoring rule (consistent scoring function) L to the forecast-observation pairs $(f_{i,t}, Y_t)_{t \in \mathbb{N}}$, that is $L_{i,t} = L(f_{i,t}, Y_t)$, $i = 1, \dots, m$, $t \in \mathbb{N}$.

Next, we clarify the notion of ‘best models’. For $i, j \in \mathcal{M}_0$ and $t \in \mathbb{N}$, consider $d_{ij,t} = L_{i,t} - L_{j,t}$ and $\mu_{ij,t} = \mathbb{E}(d_{ij,t} \mid \mathcal{F}_{t-1})$ as well as their averages $\hat{\Delta}_{ij,t} = \sum_{s=1}^t d_{ij,s}/t$ and $\Delta_{ij,t} = \sum_{s=1}^t \mu_{ij,s}/t$, where we omit the dependence of $\mu_{ij,t}$ and $\Delta_{ij,t}$ on the choice of the probability measure $\mathbb{Q}$ on $(\Omega, \mathcal{F})$ for the sake of brevity.

We follow [Giacomini and White \(2006\)](#), [Lai et al. \(2011\)](#), [Henzi and Ziegel \(2022\)](#), and [Choe and Ramdas \(2023\)](#) by defining the superior models in terms of the (average) conditional expected loss differences. In particular, we define the *strongly superior objects*

$$\mathcal{M}^{s,*} = \left\{ i \in \mathcal{M}_0 \mid \mu_{ij,t} \leq 0 \text{ a.s. for all } j \in \mathcal{M}_0, t \in \mathbb{N} \right\}, \quad (1)$$

the *uniformly weakly superior objects*

$$\mathcal{M}^{uw,*} = \left\{ i \in \mathcal{M}_0 \mid \Delta_{ij,t} \leq 0 \text{ a.s. for all } j \in \mathcal{M}_0, t \in \mathbb{N} \right\}, \quad (2)$$

and the *weakly superior objects*

$$\mathcal{M}_t^{w,*} = \left\{ i \in \mathcal{M}_0 \mid \Delta_{ij,t} \leq 0 \text{ a.s. for all } j \in \mathcal{M}_0 \right\}, \quad t \in \mathbb{N}. \quad (3)$$

For each $t \in \mathbb{N}$, we have $\mathcal{M}^{s,*} \subseteq \mathcal{M}^{uw,*} \subseteq \mathcal{M}_t^{w,*}$. Importantly, $\mathcal{M}_t^{w,*} \neq \emptyset$ for all $t \in \mathbb{N}$, whereas $\mathcal{M}^{s,*}$ and $\mathcal{M}^{uw,*}$ may be empty.

Let $(\mathcal{M}_t^*)_{t \in \mathbb{N}} \subseteq \mathcal{M}_0$ be the targeted sequence of superior objects. In the case of (1) or (2), we just consider a constant sequence. For a given confidence level $\alpha \in (0, 1)$, we call $(\widehat{\mathcal{M}}_t)_{t \in \mathbb{N}} \subseteq \mathcal{M}_0$ a *sequence of model confidence sets* or *sequential model confidence sets*, for short SMCS(s), for $(\mathcal{M}_t^*)_{t \in \mathbb{N}}$ at level α if, for any $\mathbb{Q} \in \mathfrak{B}(\Omega)$, $\mathbb{Q}(\exists t \geq 1 : \mathcal{M}_t^* \not\subseteq \widehat{\mathcal{M}}_t) \leq \alpha$, or equivalently

$$\mathbb{Q}(\forall t \geq 1 : \mathcal{M}_t^* \subseteq \widehat{\mathcal{M}}_t) \geq 1 - \alpha. \quad (4)$$

We refer to this property as a *time-uniform coverage guarantee* for the sequential model confidence sets and highlight the fact that the time quantifiers are inside the probability, which is a much stronger requirement than assuming $\mathbb{Q}(\mathcal{M}_t^* \subseteq \widehat{\mathcal{M}}_t) \geq 1 - \alpha$ for all $t \in \mathbb{N}$ (or assuming that there exists a $t \in \mathbb{N}$ with $\mathbb{Q}(\mathcal{M}_t^* \not\subseteq \widehat{\mathcal{M}}_t) \leq \alpha$); see, e.g. [Howard et al. \(2021, Section 1.1\)](#), for a historical perspective on confidence sequences.

In the paper by Hansen et al. (2011), the authors make the assumption that the (unconditional) expectation $\mathbb{E}(d_{ij,t}) = \mu_{ij}$ does not depend on t , and they define the set of superior objects as $\mathcal{M}^* = \{i \in \mathcal{M}_0 \mid \mu_{ij} \leq 0 \text{ for all } j \in \mathcal{M}_0\}$. This is in line with Diebold and Mariano (1995), who also define model (forecast) superiority with respect to the unconditional expected loss differences. For a fixed sample size, they target an asymptotic guarantee of the form $\mathbb{Q}(\mathcal{M}^* \subseteq \widehat{\mathcal{M}}) \geq 1 - \alpha$ for their model confidence set (MCS) $\widehat{\mathcal{M}}$. Our method provides a stronger guarantee but it is also the first method with finite sample validity for a fixed sample size.

Model confidence sets are also related to inference on the argmin of a vector of means, see, e.g. Futschik and Pflug (1995) and Zhang et al. (2024). In this strand of literature, the desired coverage guarantee is typically weaker in that $\mathbb{Q}(i \in \widehat{\mathcal{M}}) \geq 1 - \alpha$ should hold for each superior model i but not uniformly over all models. We discuss an adaptation of our methods to this so-called *marginal coverage* guarantee in online supplementary material D.

2.1 Discussion of the different sets of superior models

Studying different notions of superior models is important, since, depending on the particular situation, one might interpret the term ‘best’ differently. In this subsection, we discuss the hypotheses which correspond to the sequences of superior models given at (1), (2), and (3), respectively, and give concrete examples for each of them.

If it is reasonable to assume that some models outperform all other models at all time points (in terms of the conditional expected score differences), then we should try to find estimators for $\mathcal{M}^{s,*}$. We refer to this assumption as the *strong hypothesis*. The strong hypothesis particularly applies if we assume independent and identically distributed scores. Then, one typically studies the unconditional expected score differences $\mathbb{E}(d_{ij,t})$ independent of $t \in \mathbb{N}$, as it is done by Diebold and Mariano (1995) or Hansen et al. (2011), who assume stationary performances of the models. Another important instance where the strong hypothesis is reasonable is if the models have nested information sets and use them in a nearly optimal way, see Holzmann and Eulert (2014). In Section 5.1, we construct SMCSs for different Covid-19-related deaths where the assumption of nested information sets seems plausible.

On the other hand, if we assume that some models have a lower conditional expected score on average rather than at all single time points, then we should target $\mathcal{M}^{uw,*}$ under the *uniformly weak hypothesis*. Choe and Ramdas (2023, Section 4.4.) argue convincingly that there are many situations where testing for the strong hypothesis may be misleading. Inspired by their arguments, Simulation 2 below considers a forecaster which is ideal on most days and only slightly worse than the other forecasters on Sundays. Any powerful method to monitor the strongly superior models over time would exclude this forecaster from the SMCS. Nevertheless, the forecaster is superior under the uniformly weak hypothesis.

Finally, the *weak hypothesis* applies if we expect that the models evolve and their average relative performance might change over time. In Section 5.2, we study postprocessing methods for wind gust predictions. The results show that in this application it is indeed beneficial to monitor the weakly superior objects, as some models which are excluded from the SMCS at earlier time points are included again at later stages, either due to systematic changes in the underlying meteorological model, or due to the different adaptive behaviour of the methods. If we had conducted the study under any of the two stronger hypotheses, we could not have observed that some methods become again competitive towards the end of the test period.

We introduce the simulation settings of Section 4 to give concrete examples for each of the discussed hypotheses.

Simulation 1 We sample $(Y_t)_{t=1}^n$, where $Y_t \sim \mathcal{N}(Y_{t-1}, 1)$, for $t = 1, \dots, n$ and $Y_0 = 0$. We consider $m = 49$ different forecasters $\{1, \dots, m\}$ which all sequentially issue predictive distributions for $(Y_t)_{t=1}^n$ given by $f_{i,t} = \mathcal{N}(Y_{t-1} + \varepsilon_i, 1 + \delta_i)$, for $(\varepsilon_i, \delta_i) \in \{-0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6\}^2$. For these forecasters with different biases and dispersion errors, there is exactly one ideal forecaster i_0 with $\varepsilon_{i_0} = \delta_{i_0} = 0$. We assess the different forecasters with respect to the *continuous ranked probability score* (CRPS; Matheson & Winkler, 1976).

The CRPS is a proper scoring rule implying that the ideal forecaster has the lowest expected CRPS. According to the definition given at (1), this yields that $\mathcal{M}^{s,*} = \{i_0\}$. A brief introduction to (proper) scoring rules, including information about the CRPS, can be found in [online supplementary material A](#).

Simulation 2 We consider the same data-generating mechanism and the same forecasters as in Simulation 1 except that, for $t \in 7\mathbb{N}$, the forecaster i_0 now issues the forecasts $f_{i_0,t} = \mathcal{N}(Y_{t-1} + 0.3, 1.3)$. That is, forecaster f_{i_0} is still ideal on, say, weekdays and Saturdays, however, on Sundays, some forecasters have a smaller bias and dispersion error, and hence $\mathcal{M}^{s,*} = \emptyset$. However, i_0 is uniformly weakly superior, that is $\mathcal{M}^{uw,*} = \{i_0\}$.

Simulation 3 We sample i.i.d. standard normally distributed observations $(Y_t)_{t=1}^n$ and compare $m = 3$ different forecasters $i = 1, 2, 3$ which issue median predictions $m_{i,t} = Y_t + \varepsilon_{i,t}$, for $\varepsilon_{1,t} = \beta$, $\varepsilon_{2,t} = \gamma^t$, $\varepsilon_{3,t} = \delta t$, for $\beta, \delta > 0$ and $0 < \gamma < 1$. We call forecaster 1 *constantly biased*, forecaster 2 *improving* and forecaster 3 *worsening*. We assess them by

$$L(m, y) = 0.5(\Phi(m) - \Phi(y)),$$

where $m, y \in \mathbb{R}$ and Φ is the cdf of the standard normal distribution, which is a consistent scoring function for the median, see [online supplementary material A](#). If we choose the parameters as given in Section 4, we have $\mathcal{M}_t^{w,*} = \{3\}$ for $t \leq 153$, $\mathcal{M}_t^{w,*} = \{1\}$ for $153 < t < 550$, and $\mathcal{M}_t^{w,*} = \{2\}$ for $t \geq 550$. That is, whereas the worsening forecaster is initially superior, the constantly biased forecaster catches up after some time until the improving forecaster becomes the best in the end.

3 Construction of sequential model confidence sets

In this section, we provide methods to construct SMCSs for the superior models given at (1), (2), and (3). Our constructions build on e-processes, confidence sequences, and sequential multiple testing methods.

3.1 Sequential testing methods

Let $(\Omega, \mathcal{F})$ be a measurable space and $\mathfrak{B}(\Omega)$ be the family of all probability measures on $(\Omega, \mathcal{F})$. Suppose that we observe random outcomes of a process at time points $t = 1, 2, \dots$. A *statistical (null) hypothesis* $\mathcal{H} \subseteq \mathfrak{B}(\Omega)$ is a set of probability measures that are potential candidates for the true probability measure $\mathbb{P}$ governing the data generating process. For example, $\mathcal{H}$ could consist of all probability measures under which the data points are i.i.d. normally distributed with some given mean and variance.

Let $(\mathcal{F}_t)_{t \in \mathbb{N}}$ be a filtration, that is, an increasing sequence of σ -algebras on $(\Omega, \mathcal{F})$, which are all contained in $\mathcal{F}$. We interpret $\mathcal{F}_t$ as the available information at time $t \in \mathbb{N}$.

A *(super-)martingale* with respect to a probability measure $\mathbb{Q} \in \mathfrak{B}(\Omega)$ is a sequence of $\mathbb{Q}$ -integrable random variables $(X_t)_{t \in \mathbb{N}}$, which is adapted to $(\mathcal{F}_t)_{t \in \mathbb{N}}$ and satisfies the condition $\mathbb{E}_{\mathbb{Q}}[X_t | \mathcal{F}_{t-1}] = X_{t-1}$ ($\mathbb{E}_{\mathbb{Q}}[X_t | \mathcal{F}_{t-1}] \leq X_{t-1}$), for all $t \in \mathbb{N}$. We call a process $(M_t)_{t \in \mathbb{N}}$ a *test (super-)martingale* for some hypothesis $\mathcal{H} \subseteq \mathfrak{B}(\Omega)$ if it is a nonnegative (super)martingale with respect to any $\mathbb{Q} \in \mathcal{H}$, and has a starting value $M_0 \leq 1$.

Test supermartingales allow to phrase statistical testing as a betting game. Indeed, the value of the test supermartingale can be interpreted as the accumulated wealth of a gambler after having bet a number of times against the null hypothesis (Shafer, 2021). For example, in a coin-toss game with probability of heads equal to $q \in (0, 1)$, a player that bets 1\$ on heads in each round and receives 2\$ in case of success or 0\$ otherwise, cannot expect to gain money over time if the null

hypothesis $\mathcal{H}_0 : q \leq 1/2$ is true. Thus, large values of the test supermartingale give evidence against the null hypothesis.

Ville's inequality (Ville, 1939) states that any test supermartingale $(M_t)_{t \in \mathbb{N}}$ for $\mathcal{H} \subseteq \mathfrak{B}(\Omega)$ satisfies

$$\sup_{\mathbb{Q} \in \mathcal{H}} \mathbb{Q} \left(\exists t \in \mathbb{N} : M_t \geq \frac{1}{\alpha} \right) \leq \alpha, \quad \text{for all } \alpha \in (0, 1). \quad (5)$$

This lays the foundation for safe anytime-valid inference. A proof can be found in Howard et al. (2020, Section 6.1). By (5), test supermartingales yield valid sequential tests $(\psi_t)_{t \in \mathbb{N}}$ by defining $\psi_t = \mathbb{1}\{M_t \geq 1/\alpha\}$. Furthermore, Ville's inequality enables the construction of time-uniform confidence sequences (Howard et al., 2021; Ramdas et al., 2023). For $\alpha \in (0, 1)$, a $(1 - \alpha)$ -confidence sequence for a parameter sequence $(\theta_t)_{t \in \mathbb{N}}$ in some space Θ is a sequence of sets $(C_t)_{t \in \mathbb{N}}$ such that $\mathbb{Q}(\forall t \in \mathbb{N} : \theta_t \in C_t) \geq 1 - \alpha$, for all $\mathbb{Q} \in \mathfrak{B}(\Omega)$. The definition of confidence sequences resembles that of confidence sets, with the important difference that confidence sequences provide time-uniform coverage guarantees. We will use confidence sequences to construct SMCs for the sequence of weakly superior objects.

It has been shown that for some hypotheses $\mathcal{H}$, it is not possible to construct nontrivial test supermartingales (Henzi & Law, 2024; Ramdas et al., 2022). However, it may still be possible to find nontrivial e-processes. An *e-process* for some hypothesis $\mathcal{H} \subseteq \mathfrak{B}(\Omega)$ is a non-negative adapted stochastic process $(E_t)_{t \in \mathbb{N}}$ with $\mathbb{E}_{\mathbb{Q}}(E_t) \leq 1$ for all $\mathbb{Q} \in \mathcal{H}$ and all (possibly infinite) stopping times τ . Equivalently, an adapted nonnegative process $(E_t)_{t \in \mathbb{N}}$ is an e-process for $\mathcal{H}$ if, for each $\mathbb{Q} \in \mathcal{H}$, it is upper bounded by a test supermartingale. In other words, the upper bounding test supermartingale may be different for each $\mathbb{Q} \in \mathcal{H}$. Hence, any test supermartingale is also an e-process. Importantly, Ville's inequality continues to hold for e-processes (Ramdas et al., 2020).

When testing more than one hypothesis simultaneously with e-processes, multiple testing corrections are necessary. We discuss the relevant background in [online supplementary material C](#).

3.2 SMCs for the strongly superior and uniformly weakly superior objects

In this section, we construct SMCs for the strongly superior models $\mathcal{M}^{s,*}$ given at (1) and for the uniformly weakly superior models $\mathcal{M}^{uw,*}$ given at (2) using e-processes and sequential testing procedures.

We define the strong hypotheses $\mathcal{H}_{ij}^s = \{\mathbb{Q} \in \mathfrak{B}(\Omega) \mid \mu_{ij,t} \leq 0, \forall t \in \mathbb{N}\}$, and the weak hypotheses $\mathcal{H}_{ij}^{uw} = \{\mathbb{Q} \in \mathfrak{B}(\Omega) \mid \Delta_{ij,t} \leq 0, \forall t \in \mathbb{N}\}$ for $i, j \in \mathcal{M}_0$. We would like to test all these pairwise hypotheses simultaneously at each time point $t \in \mathbb{N}$. For $\bullet \in \{s, uw\}$, assume that, for any $i, j \in \mathcal{M}_0$, $(E_{ij,t})_{t \in \mathbb{N}}$ is an e-process for the hypothesis $\mathcal{H}_{ij}^\bullet$. Then, for any $i \in \mathcal{M}_0$, the arithmetic mean $E_{i,t} = 1/(m-1) \sum_{j \neq i} E_{ij,t}$ is an e-process for the intersection hypothesis $\mathcal{H}_i^\bullet = \cap_{j \neq i} \mathcal{H}_{ij}^\bullet$. We adjust the e-processes $(E_{1,t})_{t \in \mathbb{N}}, \dots, (E_{m,t})_{t \in \mathbb{N}}$ for multiple testing using the closure principle (Marcus et al., 1976) with the arithmetic mean as *e-merging* function, that is,

$$E_{i,t}^* = \min_{I \subseteq \{1, \dots, m\} : i \in I} \frac{1}{|I|} \sum_{j \in I} E_{j,t} \leq E_{i,t}, \quad i \in \mathcal{M}_0, t \in \mathbb{N}, \quad (6)$$

see [online supplementary material C](#) for further details. Importantly, each adjusted process $(E_{i,t}^*)_{t \in \mathbb{N}}$ is an e-process for the hypothesis $\mathcal{H}_i^\bullet$. As shown in Vovk and Wang (2021, Proposition 3.1.), the arithmetic mean essentially dominates all other symmetric e-merging functions. Algorithm 1 of Vovk and Wang (2021) allows for a computation of the minimum at (6) with computation time $O(m^2)$. In fact, the computation time may be improved to $O(m \log m)$, see [online supplementary material H](#) for details. Finally, for some significance level $\alpha \in (0, 1)$, we define

$$\widehat{\mathcal{M}}_t = \left\{ i \in \mathcal{M}_0 \mid E_{i,t}^* < 1/\alpha \right\}, \quad t \in \mathbb{N}. \quad (7)$$

In words, at each $t \in \mathbb{N}$, we exclude model $i \in \mathcal{M}_0$ if we may reject the hypothesis $\mathcal{H}_i^\bullet$ according to the sequential test $\psi_{i,\cdot,t}^\bullet = \mathbb{1}\{E_{i,\cdot,t}^\bullet \geq 1/\alpha\}$, and include it otherwise. The proof of the following theorem is given in [online supplementary material C](#).

Theorem 3.1 For any $\alpha \in (0, 1)$, the sequence $(\widehat{\mathcal{M}}_t)_{t \in \mathbb{N}}$ defined at (7) is an SMCS at level α for $\mathcal{M}^{\bullet,\bullet}$, $\bullet \in \{s, uw\}$, and so is its running intersection

$$\widetilde{\mathcal{M}}_t = \bigcap_{r \leq t} \widehat{\mathcal{M}}_r, \quad t \in \mathbb{N}.$$

Theorem 3.1 works since we control the family-wise error rate of the tests $\psi_{i,\cdot,t}^\bullet$ over all models i . Family-wise error rate control relies on the closure principle and some e-merging function, see [online supplementary material C](#). Alternatively, one could adjust the pairwise e-processes $E_{ij,t}$ directly to obtain adjusted e-processes for the pairwise hypotheses $\mathcal{H}_{ij}^\bullet$ and exclude the model i from the SMCS at $t \in \mathbb{N}$ if $E_{ij,t}^\bullet \geq 1/\alpha$ for some $j \neq i$. However, the corresponding tests can be shown to be uniformly less powerful, see [online supplementary material C](#).

Remark 1 Except from the family-wise error rate, the false discovery rate (FDR) is arguably the most commonly used criterion in multiple testing (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001). Recently, Wang and Ramdas (2022) studied FDR control for e-values. In principle, SMCSs could be constructed based on FDR control. However, this fundamentally changes the type of validity we are obtaining for $\widehat{\mathcal{M}}_t$. It allows us to bound the fraction of wrongly rejected models by the total number of rejected models at all time points, which is different from our target coverage guarantee at (4). We give some further details in [online supplementary material E](#).

Remark 2 Our anytime-valid methods readily extend if one is willing to weaken the simultaneous guarantee given at (4) by marginal coverage, as defined right before Section 2.1, thereby allowing to omit the adjustment step for multiple testing. The same potential gain in power may be achieved under the additional assumption that there is only one single superior object. See [online supplementary material D](#) for further details.

The following proposition provides e-processes for the strong hypotheses under the assumption of conditionally bounded loss differences. Together with Theorem 3.1, it directly leads to the derivation of SMCSs for the strongly superior objects.

Proposition 3.2 Assume that $|d_{ij,t}| \leq c_{ij,t}/2$ for $i, j \in \mathcal{M}_0$, $t \in \mathbb{N}$, and some predictable sequence $(c_{ij,t})_{t \in \mathbb{N}} \subseteq (0, \infty)$. Then, for any $i, j \in \mathcal{M}_0$, $E_{ij,t} = \prod_{r=1}^t (1 + \lambda_{ij,r} d_{ij,r})$ is an e-process for $\mathcal{H}_{ij}^s$, for any predictable $(\lambda_{ij,t})_{t \in \mathbb{N}}$ with $0 \leq \lambda_{ij,t} \leq 1/c_{ij,t}$.

Proof. The process $M_{ij,t} = \prod_{r=1}^t (1 + \lambda_{ij,r} (d_{ij,r} - \mu_{ij,r}))$ is a nonnegative martingale with $M_{ij,0} = 1$. The claim follows by observing that, for $\mathbb{Q} \in \mathcal{H}_{ij}^s$, $M_{ij,t} \geq E_{ij,t}$ for all $t \in \mathbb{N}$. $\square$

The following proposition is due to Choe and Ramdas (2023, Theorem 3) and allows to construct e-processes for the weak hypothesis under the assumption of uniformly bounded loss differences. Together with Theorem 3.1, it provides SMCSs for the uniformly weakly superior objects. In Section 3.4, we explain how we can weaken the assumption of uniformly bounded loss differences to conditionally bounded loss differences.

Proposition 3.3 Let $|d_{ij,t}| \leq c_{ij}/2$ for $i, j \in \mathcal{M}_0$, $t \in \mathbb{N}$, and some $c_{ij} > 0$. Then, for any $i, j \in \mathcal{M}_0$,

$$E_{ij,t} = \exp\left\{\lambda_{ij}t\hat{\Delta}_{ij,t} - \psi_{E,c_{ij}}(\lambda_{ij})V_{ij,t}\right\}, \quad t \in \mathbb{N},$$

is an e-process for $\mathcal{H}_{ij}^{\text{uw}}$, for any $0 \leq \lambda_{ij} \leq 1/c_{ij}$ and

$$\begin{aligned} \psi_{E,c_{ij}}(\lambda_{ij}) &= (-\log(1 - c_{ij}\lambda_{ij}) - c_{ij}\lambda_{ij})/c_{ij}^2 \quad \text{and} \\ V_{ij,t} &= \sum_{r=1}^t (d_{ij,r} - \gamma_{ij,r})^2, \end{aligned} \quad (8)$$

where $(\gamma_{ij,t})_{t \in \mathbb{N}}$ with $|\gamma_{ij,t}| \leq c_{ij}/2$ is a predictable sequence.

3.3 SMCSs for the weakly superior objects

Our SMCS construction for the weakly superior objects defined at (3) builds on time-uniform confidence regions. Proofs for this section are given in [online supplementary material F](#).

For $i, j \in \mathcal{M}_0$, $t \in \mathbb{N}$, $x \in \mathbb{R}$, let $M_{ij,t}(x)$ be such that $(M_{ij,t}(\Delta_{ij,t}))_{t \in \mathbb{N}}$ is a test supermartingale for any $\mathbb{Q} \in \mathfrak{B}(\Omega)$. Let $\mathbb{R}_0^{m \times m}$ be the family of all $m \times m$ -matrices with diagonal entries equal to zero and define for $X = (x_{ij})_{ij} \in \mathbb{R}_0^{m \times m}$,

$$M_t(X) = \frac{1}{m(m-1)} \sum_{\substack{i,j=1,\dots,m \\ i \neq j}} M_{ij,t}(x_{ij}), \quad t \in \mathbb{N}.$$

With $\Delta_t = (\Delta_{ij,t})_{ij} \in \mathbb{R}_0^{m \times m}$, also $(M_t(\Delta_t))_{t \in \mathbb{N}}$ is a test supermartingale for any $\mathbb{Q} \in \mathfrak{B}(\Omega)$. Therefore, for any $\alpha \in (0, 1)$, $C_t = \{X \in \mathbb{R}_0^{m \times m} \mid M_t(X) \leq 1/\alpha\}$ is a $(1 - \alpha)$ -confidence sequence for $(\Delta_t)_{t \in \mathbb{N}}$. Indeed, for any $\mathbb{Q} \in \mathfrak{B}(\Omega)$, by Ville's inequality,

$$\mathbb{Q}(\exists t \in \mathbb{N} : \Delta_t \notin C_t) = \mathbb{Q}(\exists t \in \mathbb{N} : M_t(\Delta_t) > 1/\alpha) \leq \alpha.$$

We construct an SMCS from C_t as follows. If there exists a $j \neq i$ such that we may reject $\Delta_{ij,t} \leq 0$, we exclude i from the SMCS. This is the case, if and only if there exists $j \neq i$ such that $C_t \cap \{X \in \mathbb{R}_0^{m \times m} \mid x_{ij} \leq 0\} = \emptyset$. Therefore, we define

$$\widehat{\mathcal{M}}_t = \{i \in \mathcal{M}_0 \mid C_t \cap \{X \in \mathbb{R}_0^{m \times m} \mid x_{ij} \leq 0\} \neq \emptyset \text{ for all } j \neq i\}, \quad t \in \mathbb{N}. \quad (9)$$

Theorem 3.4 For any $\alpha \in (0, 1)$, the sequence $(\widehat{\mathcal{M}}_t)_{t \in \mathbb{N}}$ defined at (9) is an SMCS at level α for the weakly superior objects $(\mathcal{M}_t^{\text{w},*})_{t \in \mathbb{N}}$. Its running intersection $\widetilde{\mathcal{M}}_t = \bigcap_{r \leq t} \widehat{\mathcal{M}}_r$, $t \in \mathbb{N}$ is an SMCS at level α for the uniformly weakly superior objects $\mathcal{M}^{\text{uw},*}$.

Similarly to Remark 2, we may construct more powerful SMCSs if we target a suitable marginal coverage guarantee, see [online supplementary material D](#) for details.

For computation of the SMCS, it is useful to understand the shape of the sets C_t . The following result treats a relevant case that allows for simpler computations.

Proposition 3.5 Suppose that the functions $x \mapsto M_{ij,t}(x)$ are nonnegative, convex, and decreasing for all $i, j \in \mathcal{M}_0$ and $t \in \mathbb{N}$. Then, C_t is a convex upper set, that is, for $A, B \in C_t$, it follows that $\lambda A + (1 - \lambda)B \in C_t$ for all $\lambda \in [0, 1]$, and $A \in C_t$ and $A \leq C$ implies that $C \in C_t$, where the inequality is understood componentwise.

Under the conditions of Proposition 3.5, the definition of $\widehat{\mathcal{M}}_t$ at (9) simplifies:

$$\widehat{\mathcal{M}}_t = \{i \in \mathcal{M}_0 \mid C_t \cap \{X \in \mathbb{R}_0^{m \times m} \mid x_{ij} = 0\} \neq \emptyset \text{ for all } j \neq i\}, \quad t \in \mathbb{N}. \quad (10)$$

For uniformly bounded loss differences, the following proposition provides test supermartingales that satisfy the assumptions of Proposition 3.5. It can be found in [Choe and Ramdas \(2023, Proposition 1 of Appendix\)](#). In the next section, we discuss how to accommodate conditionally bounded loss differences.

Proposition 3.6 Let $|d_{ij,t}| \leq c_{ij}/2$ for $i, j \in \mathcal{M}_0$, $t \in \mathbb{N}$, and some $c_{ij} > 0$. Then, for any $i, j \in \mathcal{M}_0$, $(M_{ij,t}(\Delta_{ij,t}))_{t \in \mathbb{N}}$ is a test supermartingale for any $\mathbb{Q} \in \mathfrak{B}(\Omega)$, where

$$M_{ij,t}(x) = \exp\left\{\lambda_{ij}t\hat{\Delta}_{ij,t} - \lambda_{ij}tx - \psi_{E,c_{ij}}(\lambda_{ij})V_{ij,t}\right\}, \quad t \in \mathbb{N},$$

for $0 \leq \lambda_{ij} \leq 1/c_{ij}$ and $\psi_{E,c_{ij}}$ and $V_{ij,t}$ are given at (8).

Since $\mathcal{M}^{uw,*} = \bigcap_{t=1}^{\infty} \mathcal{M}_t^{w,*}$, one could alternatively use the running intersection of the SMCs defined at (10) to construct SMCs for the uniformly weakly superior objects. However, the SMCs defined at (10) are much more expensive computationally. Therefore, we only compute SMCs for uniformly weakly superior objects using Theorem 3.1.

3.4 Predictable bounds and betting schemes

Our methods for the (uniformly) weakly superior objects require that the score differences are uniformly bounded over time, which is clearly more restrictive than the assumption of conditionally bounded score differences imposed for the strongly superior objects. However, one can always transform conditionally bounded score differences into uniformly bounded ones thereby modifying the underlying loss and corresponding superior objects.

Assume that $|d_{ij,t}| \leq c_{ij,t}/2$ for all $t \in \mathbb{N}$ for some predictable $(c_{ij,t})_{t \in \mathbb{N}}$. Then, the transformed loss differences $\tilde{d}_{ij,t} = d_{ij,t}/c_{ij,t}$, $t \in \mathbb{N}$ are uniformly bounded with constant $c = 1$ that is $|\tilde{d}_{ij,t}| \leq 1/2$ for all $t \in \mathbb{N}$. However, if we use the transformed loss differences, we target (possibly) different sequences of superior objects $\tilde{\mathcal{M}}^{uw,*} = \{i \in \mathcal{M}_0 \mid \tilde{\Delta}_{ij,t} \leq 0, \text{ for all } j \in \mathcal{M}_0, t \in \mathbb{N}\}$ and $\tilde{\mathcal{M}}_t^{w,*} = \{i \in \mathcal{M}_0 \mid \tilde{\Delta}_{ij,t} \leq 0, \text{ for all } j \in \mathcal{M}_0\}$ for $\tilde{\Delta}_{ij,t} = (1/t) \sum_{r=1}^t \mathbb{E}(\tilde{d}_{ij,r} \mid \mathcal{F}_{r-1}) = (1/t) \sum_{r=1}^t \mu_{ij,r}/c_{ij,r}$.

In Section 5, we use the given transformation to convert conditionally bounded CRPS differences of wind gust forecasts into uniformly bounded loss differences. Computational feasibility of this approach is due to the forecasts being either parametric or with finite support, see [online supplementary material B](#) for details. Importantly, for any proper scoring rule (consistent scoring function) S , the scaled function $\tilde{S} = S/c$ is proper (consistent) as well for any $c > 0$, an observation which justifies the transformation from a theoretical perspective. This property can be extended in the sequential setting, see [online supplementary material A](#). With the given scaling, all observations have the same maximal impact on the ranking, an effect which seems appealing for many (but surely not for all) applications.

Remark 3 For important choices of loss functions such as the quadratic or logarithmic score, the corresponding loss differences are not conditionally bounded. However, if the loss differences emerge from (or may be bounded in the tails by) a parametric family of distributions, there are other well-studied e-processes available; see, e.g. [Howard et al. \(2020\)](#). In [online supplementary material G.1](#), we provide a simulation example of mean forecasts with Gaussian errors. The resulting score differences are unbounded but still sub-exponential, which allows to use the e-processes of Proposition 3.3.

We conclude with a comment on the choice of the parameters λ_{ij} . For the uniformly weak and weak hypothesis (Propositions 3.3 and 3.6), we assume universal bounds $c_{ij} > 0$ on the pairwise loss differences, and λ_{ij} may be any fixed value in the interval $(0, c_{ij}^{-1})$. As a default, we suggest to choose $\lambda_{ij} = (2c_{ij})^{-1}$. In betting language, this corresponds to betting half of the accumulated evidence at each time step (Shafer, 2021). For the strong hypothesis (Proposition 3.2), we can choose $\lambda_{ij,t}$ predictably to increase power. There are many possible betting schemes. In Section 5.1, we propose a particular betting scheme which makes use of all previous observations and the assumption that the relative performance of two models does not change too quickly over time. Another option would be the *method of mixtures* (Robbins, 1970), where we integrate over all possible values of $\lambda_{ij,t}$ over a particular probability distribution instead of choosing one specific value. Mixtures of e-processes are again e-processes and may be expressed in closed form for some distributions (Howard et al., 2021, Appendix A.3). The method of mixtures is one of the most widely-studied techniques for constructing uniform bounds which are shown to be unimprovable under certain conditions (Howard et al., 2021; Robbins & Siegmund, 1970). However, in accordance with Waudby-Smith and Ramdas (2024), for the strong hypothesis, our experience shows that choosing $\lambda_{ij,t}$ predictably is generally more promising than the methods of mixtures. Also for the weak hypotheses, where predictable betting strategies do not apply, we could not improve the power of our methods significantly by using mixtures instead of choosing λ_{ij} fixed. For this reason, and to make computations more tractable, we suggest to work with the given e-processes directly.

4 Simulations

4.1 Simulation 1

We consider the setup given in Section 2.1. That is, we consider $n = 1,000$ realizations from a sequence $Y_t \sim \mathcal{N}(Y_{t-1}, 1)$ with $Y_0 = 0$ and assess $m = 49$ forecasters which issue sequentially probabilistic predictions with respect to the CRPS. We write $Y_t = Y_{t-1} + Z_t$, for an i.i.d. standard normally distributed sequence $(Z_t)_{t \in \mathbb{N}}$ and use the formulas provided by Grit et al. (2006) to derive

$$L_{i,t} = \sqrt{1 + \delta_i} \left(\frac{Z_t - \varepsilon_i}{\sqrt{1 + \delta_i}} \left(2\Phi \left(\frac{Z_t - \varepsilon_i}{\sqrt{1 + \delta_i}} \right) - 1 \right) + 2\phi \left(\frac{Z_t - \varepsilon_i}{\sqrt{1 + \delta_i}} \right) - \frac{1}{\sqrt{\pi}} \right), \quad (11)$$

where Φ and ϕ denote the cdf and density function of the standard normal distribution, respectively. Let $(\mathcal{F}_t)_{t \in \mathbb{N}}$ be the canonical filtration, that is, $\mathcal{F}_t$ is generated by Y_r and $f_{i,r}$ for $i = 1, \dots, m$, $r \leq t$. For any $t \in \mathbb{N}$, we have $\mathbb{E}(d_{ij,t} \mid \mathcal{F}_{t-1}) = \mathbb{E}(d_{ij,t}) = \mathbb{E}(d_{ij,1})$ by independence. Since the CRPS is a proper scoring rule, we conclude $\mathcal{M}^{s,*} = \{i \in \mathcal{M}_0 \mid \mathbb{E}(d_{ij,1}) \leq 0, \forall j \in \mathcal{M}_0\} = \{i_0\}$ for the forecaster i_0 with $\varepsilon_{i_0} = \delta_{i_0} = 0$.

We construct SMCSs for $\mathcal{M}^{s,*}$ with confidence level $\alpha = 0.1$. Upper bounds on the score differences $(d_{ij,t})_{t \in \mathbb{N}}$ are given in online supplementary material B. These do not depend on time, since biases and dispersion errors do not change, so $c_{ij,t} = c_{ij} > 0$, $t \in \mathbb{N}$. Define $\lambda_{ij,t} = (2c_{ij})^{-1}$.

Figure 1 shows the average size of the SMCS over time for $N = 1,000$ simulations. We have a 100% coverage rate for the best model i_0 , that is, in all simulation runs, the SMCS never incorrectly excludes i_0 . Hence, our methods are conservative. However, using e-processes and a sequential application of the closure principle, we cannot expect to improve the power of our methods for SMCSs, since the arithmetic average, which is the e-merging function that we used, essentially dominates any other e-merging function.

We ran Simulation 1 also for the original MCS by Hansen et al. (2011). Although its size is generally better than that of the SMCSs, the coverage rate drops below the nominal level $1 - \alpha = 0.9$ to about 0.8. The sequential implementation of the MCS is documented in online supplementary material G.2, along with another simulation in a more challenging, non-Gaussian setting. There, the coverage rates of the MCS are heavily compromised, in contrast to those of the SMCS.

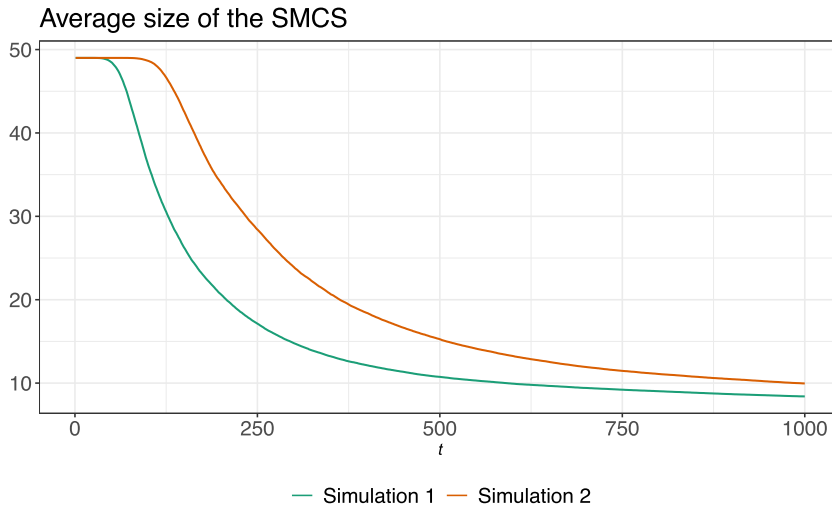


Figure 1. The average number of models in the SMCS in Simulation 1 and 2. At the end of the evaluation period, the SMCSs have an average size of 8.41 and 9.95, respectively. For both simulations, the SMCS never wrongly excludes the best model i_0 .

4.2 Simulation 2

We consider the simulation setting as given in Section 2.1 for $n = 1,000$. By (11),

$$\begin{aligned} \mathbb{E}(L_{i,t}) &= \mathbb{E}\left(Z_t \Phi\left(\frac{Z_t - \varepsilon_i}{\sqrt{1 + \delta_i}}\right)\right) - \varepsilon_i \mathbb{E}\Phi\left(\frac{Z_t - \varepsilon_i}{\sqrt{1 + \delta_i}}\right) + \varepsilon_i + 2\mathbb{E}\varphi\left(\frac{Z_t - \varepsilon_i}{\sqrt{1 + \delta_i}}\right) - \sqrt{\frac{1 + \delta_i}{\pi}} \\ &= \frac{1 + 2\sqrt{1 + \delta_i}}{\sqrt{2\pi(2 + \delta_i)}} \exp\left\{-\frac{\alpha_i^2}{2}\right\} - \varepsilon_i \Phi(-\alpha_i) + \varepsilon_i - \sqrt{\frac{1 + \delta_i}{\pi}}. \end{aligned} \quad (12)$$

For $i, j \in \mathcal{M}_0$ and $t \in \mathbb{N}$, let $\mathbb{E}(d_{ij,t} \mid \mathcal{F}_{t-1}) = \mathbb{E}(d_{ij,t}) = \mu_{ij,t}$. Since the CRPS is a proper scoring rule, we have $\mu_{i_0,t} \leq 0$ for all $t \notin 7\mathbb{N}$ and $j \in \mathcal{M}_0$. By (12), for $j \in \mathcal{M}_0$ with $(\varepsilon_j, \delta_j) \in \{-0.2, 0.2\}^2$, we have $\mu_{i_0,t} > 0$ for all $t \in 7\mathbb{N}$. Thus, $\mathcal{M}^{s,*} = \emptyset$. However, equation (12) allows to show numerically that $\Delta_{i_0,t} \leq 0$ for all $j \in \mathcal{M}_0$ and $t \in \mathbb{N}$, and thus $\mathcal{M}^{uw,*} = \{i_0\}$.

We compute SMCSs for $\mathcal{M}^{uw,*}$ with confidence level $\alpha = 0.1$ using the approach in Section 3.2. We transform the conditionally bounded CRPS differences (see [online supplementary material B](#)) as proposed in Section 3.4 into uniformly bounded score differences $|\tilde{d}_{ij,t}| \leq c/2$ for $c = 2$, and let $\lambda_{ij} = (2c)^{-1} = 1/4$. [Figure 1](#) displays the average SMCS size across time, which declines slightly slower than in Simulation 1. This is to be expected, since we work with a weaker hypothesis. Regarding coverage of the superior object, the SMCS is still conservative and exhibits a coverage rate of 100%.

It could be interesting to monitor the e-processes that we use to collect evidence on predictive performance over time. Due to the large number of models considered here, this is not so practical, but we illustrate this idea in [online supplementary material G.1](#), where only nine models are considered in the experiment.

In [online supplementary material D](#), we illustrate the potential gain in power by replacing simultaneous coverage by marginal coverage (see Remark 2) by comparing the average size of the SMCS in Simulations 1 and 2 with and without multiple testing correction.

4.3 Simulation 3

We consider the simulation setting from Section 2.1. Let $i = 1, 2, 3$ be forecasters issuing median predictions $m_{i,t} = Y_t + \varepsilon_{i,t}$, $t \in \mathbb{N}$, for i.i.d. standard normally distributed $(Y_t)_{t \in \mathbb{N}}$ and

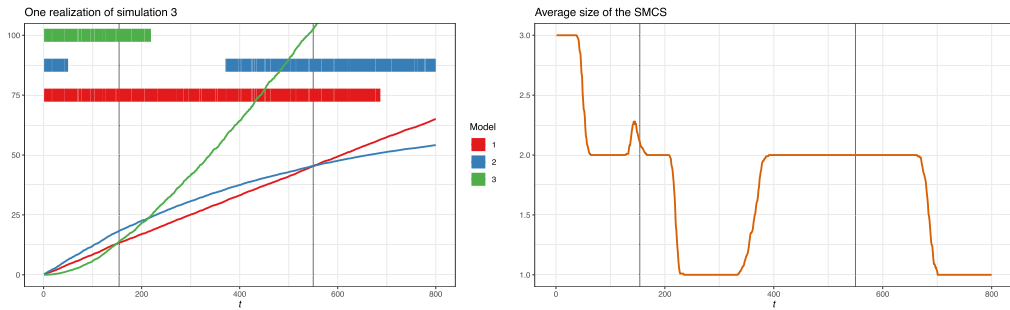


Figure 2. Left: Realized accumulated losses $\sum_{r=1}^t L_{i,r}$ for one realization: Worsening forecaster, improving forecaster, constantly biased forecaster. The black vertical lines indicate $t = 154$ and $t = 550$. The resulting SMCS is given in the upper part with the respective colours. Right: Average size of the SMCS over $N = 100$ realizations.

$\varepsilon_{1,t} = \beta$, $\varepsilon_{2,t} = \gamma^t$, $\varepsilon_{3,t} = \delta t$, for $\beta, \delta > 0$ and $0 < \gamma < 1$. For the loss $L(m, y) = 0.5|\Phi(m) - \Phi(y)|$, we get by independence that $2\mu_{ij,t} = \mathbb{E}(\Phi(m_{i,t})) - \mathbb{E}(\Phi(m_{j,t})) = \Phi(\varepsilon_{i,t}/\sqrt{2}) - \Phi(\varepsilon_{j,t}/\sqrt{2})$ for $t \in \mathbb{N}$, $i, j = 1, 2, 3$, thus

$$\mathcal{M}_t^{w,*} = \left\{ i \mid \sum_{r=1}^t \Phi\left(\frac{\varepsilon_{i,r}}{\sqrt{2}}\right) \leq \sum_{r=1}^t \Phi\left(\frac{\varepsilon_{j,r}}{\sqrt{2}}\right), \text{ for all } j \right\}. \quad (13)$$

We let $\beta = 0.6$, $\gamma = 0.998$, and $\delta = 0.008$. Then, one can numerically show that $\mathcal{M}_t^{w,*} = \{3\}$ for $t \leq 153$, $\mathcal{M}_t^{w,*} = \{1\}$ for $153 < t < 550$, and $\mathcal{M}_t^{w,*} = \{2\}$ for $t \geq 550$.

We sample $n = 800$ observations and construct the SMCSs defined in Section 3.3 for $(\mathcal{M}_t^{w,*})_{t \leq n}$, at level $\alpha = 0.1$. We use the fact that $(d_{ij,t})_{t \in \mathbb{N}}$ is uniformly bounded with $c = 1$ and let $\lambda_{ij} = 1.1^{-1}$, $i, j = 1, 2, 3$. Figure 2 shows the accumulated observed losses $\sum_{r=1}^t L_{i,r}$ and the resulting SMCS for one realization. The SMCS correctly excludes the improving forecaster at the beginning and the worsening forecaster relatively quickly after the first change point ($t = 154$) until it includes the improving forecaster again and finally excludes the constantly biased forecaster. The average size of the SMCS for $N = 100$ runs varies between one and two models after a short initial period with three models. The SMCS includes the superior objects given at (13) at all time points.

Again, we show in [online supplementary material D](#) how one can reduce the average size of the SMCS under a weaker coverage guarantee.

5 Case studies

5.1 Covid-19 case study

After the outbreak of the Covid-19 pandemic, the Reich lab at the University of Massachusetts collaborated with the United States centres for disease control and prevention (CDC) to create the Covid-19 forecast hub, a repository containing point and probabilistic forecasts for incident cases, hospitalizations, and Covid-19-related deaths. The repository was founded in March 2020, and attracted submissions of forecasts from more than 60 different teams. The collection of forecasts is still ongoing, although most of the models had stopped submitting forecasts by January 2024, due to the subsidence of the pandemic.

5.1.1 Data

The dataset we use in this application, is discussed in detail in [Cramer, Huang, et al. \(2022\)](#) and [Cramer, Ray, et al. \(2022\)](#), and publicly available at the Covid-19 forecast hub GitHub repository and on the Zoltar forecast archive. It contains forecasts with corresponding observations for key epidemiology indices for 55 different locations in the U.S. as well as on the aggregated U.S. national level. The forecasts are issued by a total of 69 different forecast models and reported at 23 different quantile levels with forecast horizons ranging from 1 day to 20 weeks. Some models are not

statistically comparable since they have consistently issued their forecasts on different weekdays. There are also models that started issuing forecasts late into the forecasting period and others that stop early. Finally, there are models with a large number of missing values. In principle, one could use missing value imputation, or assign a default score to missing values so that the number of comparable models is higher. However, the purpose of this study is to demonstrate how sequential model confidence sets work, so we do not go further in this direction. We only include models that are directly comparable without any preprocessing.

Specifically, we consider 1-week-ahead forecasts for Covid-related deaths on the national level, issued weekly in the period from 06 June 2020 to 04 March 2024, and focus on the comparison of the following $m = 6$ different forecasting models: Firstly, we consider the benchmark model *baseline*, which naively issues the most recent outcome as the median prediction for the following week and forms a predictive distribution around this median prediction based on the past weekly incidences. Secondly, we consider two epidemic models *PSI-draft* and *MOBS-gleam* issuing their forecasts based on epidemiological assumptions, and the neural network based model *GT-deep*, which was the first purely data-driven model to be included in the Covid-19 forecast hub. Finally, we consider the summary models *ensemble* and *CDC-ensemble*, where the former combines the most recent predictions of all other models into one predictive distribution, and the latter is supposed to improve the former by only considering the 10 current best models, measured by the average weighted interval score over the 12 most recent weeks. For a more detailed documentation of the models, we refer to [Cramer, Huang, et al. \(2022\)](#).

We perform our comparison under the strong hypothesis, since we expect the ensemble models to have consistent superior performance over the individual models, which in turn are expected to be better than the naive baseline model issuing its forecasts only based on the previous outcomes and on no additional information. Furthermore, we can argue that in such a highly important prediction task, we should seek to detect the models that issue accurate predictions consistently over time and not just on average.

The outcomes as well as the forecasts exhibit unstable behaviour and large variation in scale, see e.g. the upper left panel of [Figure 3](#), which suggests to apply a log transformation on all quantities of interest. Moreover, forecast differences are typically not distinctive for the median, while they become more apparent when we look at the tails, see the lower panels in [Figure 3](#).

5.1.2 Methods and implementation details

We assess the different quantile forecasts at level $\tau \in (0, 1)$ with respect to the generalized piecewise-linear quantile scoring function $S_\tau(x, y) = (1\{x \geq y\} - \tau)(\log(x) - \log(y))$ for $x, y \in \mathbb{R}$, see [Gneiting \(2011\)](#) and [online supplementary material A](#). In [online supplementary material B](#), it is shown that the resulting quantile score differences are conditionally bounded. More precisely, for τ -quantile predictions $(x_{i,t})_{t \in \mathbb{N}}, (x_{j,t})_{t \in \mathbb{N}} \subseteq \mathbb{R}$ issued by the models $i, j = 1, \dots, m$, we have $|d_{ij,t}| = |S_\tau(x_{i,t}, y_t) - S_\tau(x_{j,t}, y_t)| \leq c_{ij,t}/2$ for

$$c_{ij,t} = 2 \max\{\tau, 1 - \tau\} |\log(x_{i,t}) - \log(x_{j,t})|, \quad t \in \mathbb{N}. \quad (14)$$

We apply our methods from [Section 3.2](#) to construct SMCs for the strongly superior objects, using the e-processes $E_{ij,t} = \prod_{r=1}^t (1 + \lambda_{ij,r} d_{ij,r})$ for a predictable sequence $(\lambda_{ij,t})_{t \in \mathbb{N}}$ with $0 < \lambda_{ij,t} < c_{ij,t}^{-1}$, $t \in \mathbb{N}$. For $i, j = 1, \dots, m$ and $t \in \mathbb{N}$, we suggest $\lambda_{ij,t} = (K_{ij,t} c_{ij,t} + \varepsilon)^{-1}$ for some small $\varepsilon > 0$, included as a safeguard for the scenario $c_{ij,t} = 0$, which would be the case if the two forecasters issued the same prediction, and

$$K_{ij,t} = \frac{2 - \left| \tau - \frac{1}{2} \right|}{1 + \left| \tau - \frac{1}{2} \right|} \cdot \frac{3\pi/2 + \arctan(-d_{ij,t-1})}{\pi} \geq 1.$$

The coefficient $K_{ij,t}$ depends on the sign and magnitude of the most recent score difference $d_{ij,t-1}$ as well as of the centrality of τ : If $d_{ij,t-1} > 0$, then we have evidence against the null hypothesis that

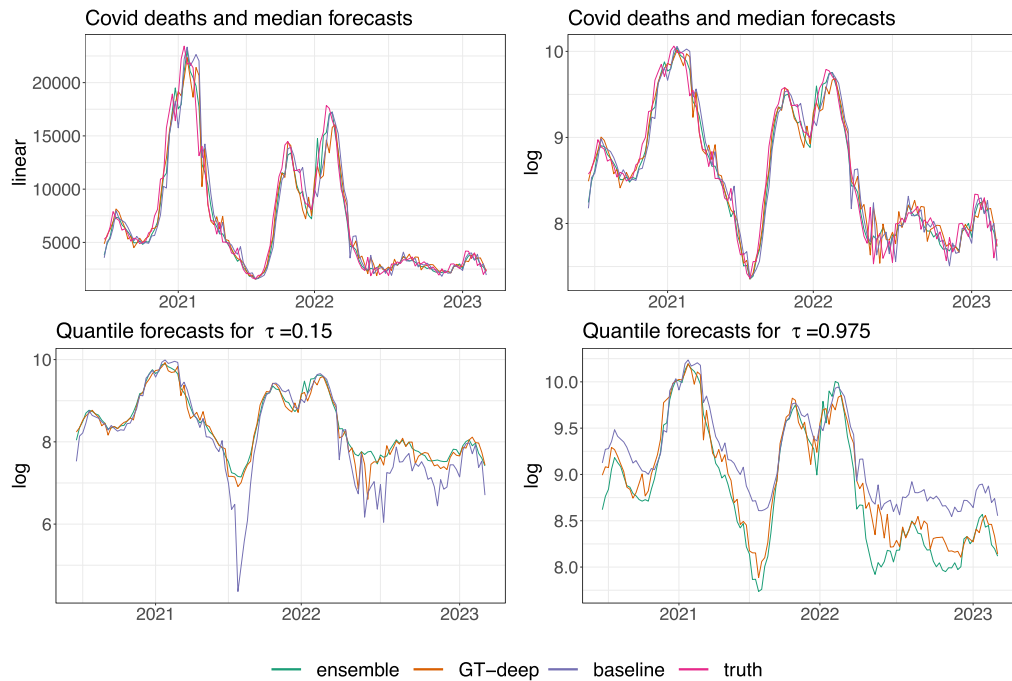


Figure 3. Time progression of predicted and actual Covid-related mortality in linear (upper left) and log scale (upper right). For the forecasts of the tails, the differences between the different models are more pronounced than for the median, see the lower panels for $\tau = 0.15, 0.975$.

model i is strictly superior than model j , and the quantity $K_{ij,t}$ becomes small. That is, the parameter $\lambda_{ij,t}$ tends to be larger, and we bet more against the null hypothesis. Finally, the primary multiplicative factor, which takes values in $[1, 2]$, is used to reduce the influence of the term $\max\{\tau, 1 - \tau\}$ on the bound in (14).

5.1.3 Results

Figure 4 shows the resulting SMCSs for four selected quantile levels with a confidence level of $\alpha = 0.1$. As a first observation, we see that the power of the SMCS varies with the choice of the quantile level τ : Whereas the SMCS detects three deficient models for $\tau = 0.1, 0.3, 0.5$, it only has excluded two models by the end of the observation period for $\tau = 0.7$. For $\tau = 0.1$, the SMCS needs considerably more observations to reject the three models, which are excluded much faster at the median. The lower power in the tails of the distribution may be due to the factor $\max\{\tau, 1 - \tau\}$ in the bound (14). Overall, the results show that the two ensemble models perform best, as they are included in the SMCS across the entire time period, and at all quantile levels. The superior performance of the two ensemble models was expected, since they combine the forecasts of all other models and thus have access to a larger information set. Interestingly, the SMCS does not detect any significant difference between ensemble and CDC-ensemble, where the latter was supposed to improve the former by combining only the best-performing forecasting models. The naive baseline performs surprisingly well and even outperforms the epidemiological model, PSI-draft, which is excluded earlier than the baseline for all given values of τ . We may conclude that PSI-draft is not competitive against the other models and should therefore not be considered as an accurate forecasting model for the pandemic. Finally, the second epidemiological model, MOBS-gleam, performs almost always better than PSI-draft but not as good as the neural network based GT-deep.

To conclude, we highlight once again the key feature that SMCSs allow us to collect statistical evidence sequentially over time. Hence, if we were back in the pandemic, by using SMCSs, we

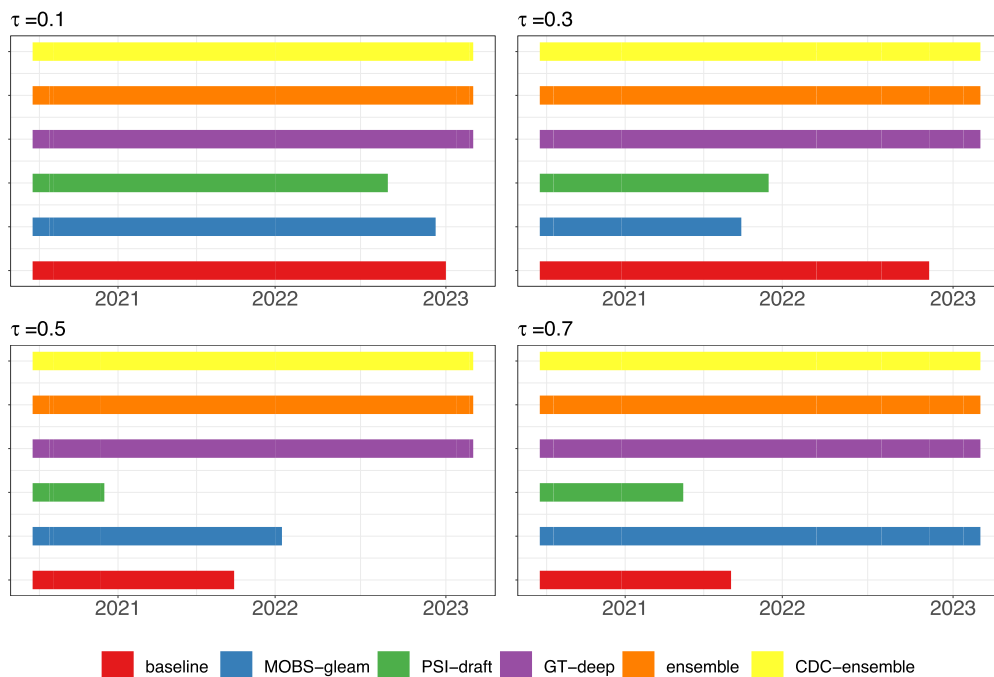


Figure 4. SMCSs at four selected quantile levels with confidence level $\alpha = 0.1$.

could monitor the models' performances from the beginning on and would not have to wait until we have collected enough observations to perform a statistical test for predictive ability. Quickly, we would reject PSI-draft as an accurate model and not consider it in our analysis and policy-making anymore. Also MOBS-gleam would probably have been rejected as a competitive model by approximately the end of 2021. Clearly, in such emergency situations, it has already been common practice to assess the forecasting models on a regular basis. However, this practice may lack theoretical justification, and our methodology provides statistically safe methods to do so.

We focused primarily on the sequential monitoring of model performances but our methodology also allows for optional stopping. For example, a possible stopping criterion back in the pandemic could have been the first day when there are only half of the original models left.

Throughout the article, we have used a significance level equal to $\alpha = 0.1$, but clearly there is freedom to do otherwise. The choice of α affects the interpretation of the SMCS. For example, when $\alpha = 0.1$, we can deduce that the SMCS contains the optimal Covid-19 forecasting model with 90% confidence.

The choice of α also affects the size of the SMCS. Larger values of α would lead to earlier detections and to a smaller SMCS. However, they would also translate to a lower level of confidence. In practice, we should keep this tradeoff in mind when choosing α . Our goal to narrow down the original set of models to a small SMCS competes with the high level of confidence that is necessary in various applications.

5.2 Postprocessing of wind gust predictions

Weather services use numerical weather prediction (NWP) models for predicting the future state of the atmosphere. These NWP models quantify forecast uncertainty using ensemble predictions, where each ensemble member represents a different future scenario. However, such ensemble forecasts are typically subject to systematic biases and dispersion errors. Hence, the forecasts ought to be statistically postprocessed to generate accurate and reliable predictions (Gneiting & Raftery, 2005; Vannitsem et al., 2018).

Schulz and Lerch (2022) present a comprehensive comparison of multiple approaches for statistical postprocessing of ensemble forecasts for wind gusts. Here, we will build on their case study and compare the different forecasting models for probabilistic wind gust prediction by using SMCSs under the weak hypothesis.

5.2.1 Data

We use the dataset in Primo et al. (2023), which is an extension of the original data in Schulz and Lerch (2022). The forecast data is based on an operational NWP model from the German weather service [Deutscher Wetterdienst (DWD)] in the period from 08 December 2010 to 30 June 2023 with corresponding observations at 174 geographically diverse weather stations in Germany. The forecasts are initialized at 00 UTC and have a forecast horizon of up to 21 hr. Here, we consider only those forecasts with a lead time of 18 hr referring to 18 UTC and the $K = 166$ stations with at least 2,600 observations.

Next to the ensemble forecasts (EPS), we include the eight postprocessing methods applied by Schulz and Lerch (2022). Statistical postprocessing methods are typically distributional regression models that use ensemble predictions from the NWP model as input data. The methods we compare can be divided into three groups: Basic techniques that only make use of the underlying wind gust ensemble predictions (EMOS, MBM, and IDR), established machine learning methods for postprocessing that incorporate additional covariates (EMOS-GB and QRF), and neural network-based approaches (DRN, BQN, and HEN). While the first two groups fit a separate model for each station, the neural networks estimate one locally adaptive model for all stations. Further, the models differ in the resulting forecast distributions, which range from parametric (EMOS, EMOS-GB, and DRN) to semi- and nonparametric types (all others). The postprocessing models are trained on the period from 2010 to 2015 using the same configuration as in Schulz and Lerch (2022) and evaluated on the remaining period from 2016 to 2023.

To improve the forecast quality, NWP models are continuously developed further and frequently updated. Within the period from 2010 to 2023, the underlying NWP model has undergone several updates, of which three change the systematic errors of the ensemble forecasts drastically (on 22 March 2017, 16 May 2018, and 10 February 2021). These model updates present a challenge for statistical postprocessing methods, as the corrections learned on previous NWP model versions may not lead to the same improvements when applied to the current model version. In essence, the training data become less representative of the test data after each update. The methods considered here have been trained on data until the end of 2015, which is before the first major model update in 2017. Still, they are applied to ensemble forecasts that have been generated by another NWP model version, which has undergone up to three major updates. Thus, we expect the behaviour both of the ensemble predictions and the postprocessed forecasts to change systematically over time. For further details on the data and postprocessing methods, we refer to Schulz and Lerch (2022) and Primo et al. (2023).

5.2.2 Methods and implementation details

We assess the $m = 9$ different forecasting models with the CRPS and construct SMCSs for the weak hypothesis as proposed in Section 3.3. Since the forecasts are predictable, the CRPS differences are conditionally bounded, see online supplementary material B. We use the transformation we discussed in Section 3.4 to obtain uniformly bounded score differences. That is, $|d_{ij,t}^k| \leq c/2$ for the resulting CRPS differences at station $k \leq K$, for $c = 2$, $i, j \leq m$, $t \in \mathbb{N}$. We set $\lambda_{ij,t} = (2c)^{-1} = 1/4$ and $\alpha = 0.1$. For inference on all stations simultaneously, we consider the average scaled CRPS differences $d_{ij,t} = \sum_{k=1}^K d_{ij,t}^k / K$, for $i, j \leq 9$ and $t \in \mathbb{N}$.

5.2.3 Results

First, for the SMCS with respect to the loss differences averaged over all stations, our conclusions regarding the performance of postprocessing methods align with the results in Schulz and Lerch (2022). The left panel of Figure 5 displays the SMCS and shows that the two neural network-based approaches, DRN and BQN, which performed best in the former study, are included in the SMCS for the entire time period. The other models fall out of the SMCS over time, where the order

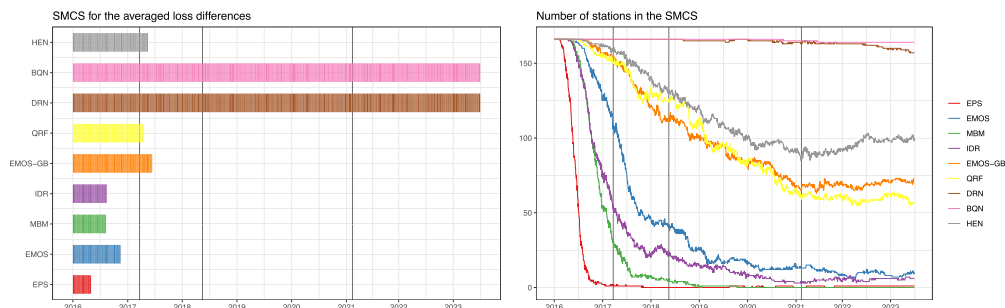


Figure 5. SMCS of averaged loss differences over all stations (left) and the number of stations where a method is included in the SMCS dependent on time (right). The vertical black lines indicate the three major NWP model updates.

coincides with that found in the former study, e.g. the EPS is omitted first. In this case, where we average the performances over all stations, the SMCS behaves as for the stronger hypotheses, that is, once we eliminate a model, we do not include it anymore.

Second, we consider the SMCS station-wise. The right panel of Figure 5 gives a summary, as it shows the number of stations for which a forecasting model is included in the corresponding SMCS. The EPS is excluded from most SMCSs already within the first year, the basic techniques EMOS, MBM, and IDR are also excluded relatively fast, and that the machine learning approaches EMOS-GB and QRF fall out of the SMCSs successively over time, until early 2021 when the last NWP model update occurred. Interestingly, the number of stations that include the HEN network approach increases after the last NWP model update, a behaviour only detectable using the weak hypothesis. Finally, the BQN and DRN approach are included at almost all stations over the entire period, only with DRN being excluded over the last year at some stations.

Taking a closer look at the individual stations, we observe different kinds of local behaviour that is not visible when averaging over the loss differences. We can broadly divide the stations into three categories: Regular (56 stations), dynamic (55), and fluctuating (55). Figure 6 shows one exemplary station for each of the three categories as well as the mean size of the station-wise SMCSs over time. The SMCS of a *regular* station behaves (almost) equally as for the stronger hypotheses, which is also seen in the plot of the mean SMCS size over the time period. The *dynamic* stations show a different pattern. As for the regular stations, some of the inferior models are excluded after a certain time period. However, in contrast to the regular stations, some methods that have been eliminated before are now included again, coming closer to the end of the time period. This behaviour becomes apparent when looking at the mean size of the SMCSs, which increases after the last update of the underlying NWP model, which systematically changes the predictive performance of the methods. These cases are interesting, since any test that would be performed just at the end of the evaluation period (or at some given time point in the middle) may not detect the deficiencies of these models. Consider, for example, the SMCS at station 10044, Leuchtturm Kiel, displayed at the top right of Figure 6, where EMOS, IDR, and HEN are excluded from the SMCS for long periods and become competitive again later on. Although these three models do not perform as strongly as DRN and BQN, the SMCS indicates that they are still better than the remaining models for this station. Finally, we have the *fluctuating* stations, where at least one model is repeatedly excluded and included from the SMCS. As for the other two groups, this behaviour is reflected in the mean size of the SMCS, which is oscillating more for the fluctuating stations than the other two categories. In all cases, the sequential nature of the SMCS yields a deeper insight into the performance of the methods than tests for predictive ability which are issued at single time points only.

6 Discussion

We introduce sequential model confidence sets and provide methodology to construct them with respect to three important notions of forecast superiority. SMCSs allow to continuously monitor

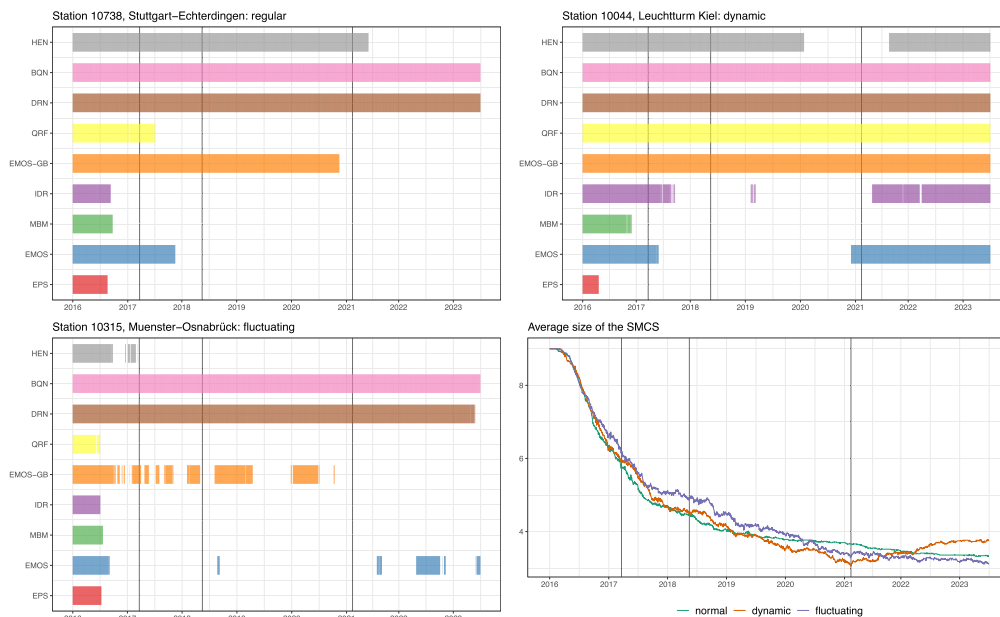


Figure 6. SMCSs for one exemplary station in each of the three categories (top left: normal; top right: dynamic; bottom left: fluctuating) and the mean size of the SMCS averaged over the stations of each category dependent on the time (bottom right).

the performance of some given statistical models. They incorporate the uncertainty in choosing models by being flexible in size, rely on minimal distributional assumptions, and achieve anytime-valid coverage guarantees.

We have provided coverage guarantees of all superior models with high probability uniformly over time. Other coverage guarantees could be of interest but require somewhat different approaches. The literature on inference on the argmin of a vector of means typically focuses on the guarantee that each superior model is covered with high probability. Future work could consider time-uniform extensions of such guarantees. Furthermore, one could aim for a time-uniform bound of the expected fraction of wrongly rejected models by the total number of rejected models. This can be achieved by time-uniform control of the FDR, and we give first results in this direction in [online supplementary material E](#).

We have mainly focused on (conditionally) bounded score differences. However, important loss functions such as the squared error or the logarithmic score do not satisfy this requirement. Nevertheless, our general theory is not limited to bounded score differences. We have provided a simulation study for SMCSs with the squared error loss in [online supplementary material G. 1](#). Nonetheless, further research is warranted to extend the understanding and implementation of SMCSs to encompass more general loss functions. Finally, it would be interesting to examine how our results extend to important information criteria, and to apply sequential model confidence sets for sequential model selection.

Acknowledgments

The comments of three anonymous referees significantly improved the quality of the article. We thank Lutz Dümbgen, Peter Grünwald, and Tim Stephan for valuable discussions and inputs. We thank Cristina Primo, Reinhold Hess, and Sebastian Trepte for providing the forecast and observation data, and Robert Redl for assistance in data handling.

Conflicts of interest: All authors declare no conflicts of interest.

Funding

S.A. and J.Z. gratefully acknowledge financial support from the Swiss National Science Foundation under grant number 200021_192269. B.S. gratefully acknowledges funding within the project C5 “Dynamical feature-based ensemble postprocessing of wind gusts within European winter storms” of the Transregional Collaborative Research Center SFB/TRR 165 “Waves to Weather” funded by the German Research Foundation (DFG).

Data availability

Code in R for the case studies and replication material for the simulations are available at <https://github.com/GGavrilos/SMCS>. Computations for the simulations and parts of the wind gust case study have been performed on UBELIX (<https://ubelix.unibe.ch/>), the HPC cluster of the University of Bern.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

References

- Benjamini Y., & Hochberg Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 57, 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Benjamini Y., & Yekutieli D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29, 1165–1188. <https://doi.org/10.1214/aos/1013699998>
- Choe Y. J., & Ramdas A. (2023). Comparing sequential forecasters. *Operations Research*, 72, 1368–1387. <https://doi.org/10.1287/opre.2021.0792>
- Cramer E. Y., Huang Y., Wang Y., Ray E. L., Cornell M., Bracher J., Brennen A., Rivadeneira A. J. C., Gerding A., House K., Jayawardena D., Kanji A. H., Khandelwal A., Le K., Mody V., Mody V., Niemi J., Stark A., Shah A., ... Reich N. G. (2022). The United States covid-19 forecast hub dataset. *Scientific Data*, 9(1), 462. <https://doi.org/10.1038/s41597-022-01517-w>
- Cramer E. Y., Ray E. L., Lopez V. K., Bracher J., Brennen A., Castro Rivadeneira A. J., Gerding A., Gneiting T., House K. H., Huang Y., & Jayawardena D. (2022). 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*, 119(15), Article e2113561119. <https://doi.org/10.1073/pnas.2113561119>
- Diebold F., & Mariano R. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13, 253–263. <https://doi.org/10.1080/07350015.1995.10524599>
- Futschik A., & Pflug G. (1995). Confidence sets for discrete stochastic optimization. *Annals of Operations Research*, 56, 95–108. <https://doi.org/10.1007/BF02031702>
- Giacomini R., & White H. (2006). Tests of conditional predictive ability. *Econometrica: Journal of the Econometric Society*, 74(6), 1545–1578. <https://doi.org/10.1111/j.1468-0262.2006.00718.x>
- Gneiting T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106, 746–762. <https://doi.org/10.1198/jasa.2011.r10138>
- Gneiting T., & Raftery A. E. (2005). Weather forecasting with ensemble methods. *Science*, 310, 248–249. <https://doi.org/10.1126/science.1115255>
- Grimt E. P., Gneiting T., Berrocal V. J., & Johnson N. A. (2006). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, 132, 2925–2942. <https://doi.org/10.1256/qj.05.235>
- Grünwald P., de Heide R., & Koolen W. (2024). Safe testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 86(5), 1091–1128. <https://doi.org/10.1093/jrsssb/qkae011>
- Hansen P. R., Lunde A., & Nason J. M. (2011). The model confidence set. *Econometrica: Journal of the Econometric Society*, 79(2), 453–497. <https://doi.org/10.3982/ECTA5771>
- Henzi A., & Law M. (2024). A rank-based sequential test of independence. *Biometrika*, 111(4), 1169–1186. <https://doi.org/10.1093/biomet/asae023>
- Henzi A., & Ziegel J. F. (2022). Valid sequential inference on probability forecast performance. *Biometrika*, 109(3), 647–663. <https://doi.org/10.1093/biomet/asab047>
- Holzmann H., & Eulert M. (2014). The role of the information set for forecasting—with applications to risk management. *Annals of Applied Statistics*, 8(1), 595–621. <https://doi.org/10.1214/13-AOAS709>
- Howard S. R., Ramdas A., McAuliffe J., & Sekhon J. (2020). Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17, 257–317. <https://doi.org/10.1214/18-PS321>

- Howard S. R., Ramdas A., McAuliffe J., & Sekhon J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *Annals of Statistics*, 49(2), 1055–1080. <https://doi.org/10.1214/20-AOS1991>
- Lai T. L., Gross S. T., & Shen D. B. (2011). Evaluating probability forecasts. *Annals of Statistics*, 39(5), 2356–2382. <https://doi.org/10.1214/11-AOS902>
- Marcus R., Peritz E., & Gabriel K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3), 655–660. <https://doi.org/10.1093/biomet/63.3.655>
- Matheson J. E., & Winkler R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22, 1087–1096. <https://doi.org/10.1287/mnsc.22.10.1087>
- Nan F. (2009). *Forecasting next-day electricity prices: From different models to combination* [PhD thesis]. University of Padova.
- Peng Y., Albuquerque P. H. M., de Sá J. M. C., Padula A. J. A., & Montenegro M. R. (2018). The best of two worlds: Forecasting high frequency volatility for cryptocurrencies and traditional currencies with support vector regression. *Expert Systems with Applications*, 97, 177–192. <https://doi.org/10.1016/j.eswa.2017.12.004>
- Primo C., Schulz B., Lerch S., & Hess R. (2023). *Comparison of model output statistics and neural networks to postprocess wind gusts*. In *Symposium on applications of mathematical sciences* (pp. 153–180). Springer.
- Ramdas A., Grünwald P., Vovk V., & Shafer G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4), 576–601. <https://doi.org/10.1214/23-STS894>
- Ramdas A., Ruf J., Larsson M., & Koolen W. (2020). ‘Admissible anytime-valid sequential inference must rely on nonnegative martingales’, arXiv, arXiv:2009.03167v3, preprint: not peer reviewed.
- Ramdas A., Ruf J., Larsson M., & Koolen W. M. (2022). Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 141, 83–109. <https://doi.org/10.1016/j.ijar.2021.06.017>
- Robbins H. (1970). Statistical methods related to the law of the iterated logarithm. *Annals of Mathematical Statistics*, 41(5), 1397–1409. <https://doi.org/10.1214/aoms/1177696786>
- Robbins H., & Siegmund D. (1970). Boundary crossing probabilities for the wiener process and sample sums. *Annals of Mathematical Statistics*, 41(5), 1410–1429. <https://doi.org/10.1214/aoms/1177696787>
- Schulz B., & Lerch S. (2022). Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Monthly Weather Review*, 150(1), 235–257. <https://doi.org/10.1175/MWR-D-21-0150.1>
- Shafer G. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 184, 407–431. <https://doi.org/10.1111/rssa.12647>
- Vannitsem S., Wilks D. S., & Messner J. (Eds.). (2018). *Statistical postprocessing of ensemble forecasts*. Elsevier.
- Ville J. (1939). *Étude Critique de la Notion de Collectif* [PhD thesis]. L’Université de Paris.
- Vovk V., & Wang R. (2021). E-values: Calibration, combination and applications. *Annals of Statistics*, 49(3), 1736–1754. <https://doi.org/10.1214/20-AOS2020>
- Wang R., & Ramdas A. (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 84(3), 822–852. <https://doi.org/10.1111/rssb.12489>
- Waudby-Smith I., & Ramdas A. (2024). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 86(1), 1–27. <https://doi.org/10.1093/jrssb/qkad009>
- Zhang T., Lee H., & Lei J. (2024). ‘Winners with confidence: Discrete argmin inference with an application to model selection’, arXiv, arXiv:2408.02060v4, preprint: not peer reviewed.