Combining forecasts using stochastic dominance *

Mehmet Pinar[†] Edge Hill University

Thanasis Stengos[‡] University of Guelph

M. Ege Yazgan[§] Istanbul Bilgi University

April 28, 2014

Abstract

This paper derives optimal (worst) forecast combinations based on stochastic dominance efficiency (SDE) analysis with differential forecast weights. For the optimal (worst) forecast combination, SDE will minimize (maximize) the cumulative density functions (cdfs) of the forecast errors at different parts of the forecast error distribution by combining different time-series model-based forecasts. Using two exchange rate series on weekly data for the Japanese Yen/U.S. Dollar and U.S. Dollar/Great Britain Pound for the period from 1975 to 2010, we find that the optimal forecast combinations with SDE weights perform better than different forecast selection and combination methods for the majority of the cases. However, there are also some very few cases where some other forecast selection and combination model performs equally well or even better at some part of the forecast error distribution. We also find that the worst forecast combination with SDE weights is the worst one at all parts of the forecast error distribution when compared with all other forecast selection and combination models. In that context, the random walk model is the model that consistently contributes with considerably more than equal weight to the worst forecast combination for all variables being forecasted and for all forecast horizons at all parts of the forecast error distribution. For the optimal forecast combination, different models work better at different areas of the empirical distribution, which also includes the average forecast combination. Among all models, the flexible neural network autoregressive model and the self-exciting threshold autoregressive model contribute the most to the optimal forecast combination, with weights that are considerably greater than the equal ones.

JEL Classifications: C12; C13; C14; C15; G01

Key Words: Nonparametric Stochastic Dominance; Mixed Integer Programming; Forecast Combinations

*Thanasis Stengos would like to acknowledge financial support from The Scientific and Technological Research Council of Turkey (TUBİTAK).

[†]mehmet.pinar@edgehill.ac.uk

[‡]tstengos@uoguelph.ca

[§]eyazgan@bilgi.edu.tr

1 Introduction

Since the seminal work of Bates and Granger (1969), combining the forecasts of different models, rather than relying on the forecasts of individual models, has come to be viewed as an effective way to improve the accuracy of predictions regarding a certain target variable. A significant number of theoretical and empirical studies, e.g., Timmermann (2006) and Stock and Watson (2004), have been able to demonstrate the superiority of combined forecasts over single-model-based predictions.

In this context, the central question is to determine the optimal weights used in the calculation of combined forecasts. In combined forecasts, the weights attributed to each model depend on the model's out of sample performance. Over time, the forecast errors used for the calculation of optimal weights change; thus, the weights themselves vary over time. However, in empirical applications, numerous papers (Clemen, 1989; Stock and Watson, 1999a, 1999b, 2004; Hendry and Clements, 2004; Smith and Wallis, 2009; Huang and Lee, 2010; Aiolfi et al., 2011; Geweke and Amisano, 2012) have found that equally weighted forecast combinations often outperform or perform almost as well as estimated optimal forecast combinations. This finding is frequently referred as the "forecast combination puzzle" by Stock and Watson (2004) because the efficiency cost of estimating the additional parameters of an optimal combination exceeds the variance reduction gained by deviating from equal weights.[1] Overall, even though different optimal forecast combination weights are derived for static, dynamic, or time-varying situations, most empirical findings suggest that the simple average forecast combination outperforms more sophisticated weighting schemes.

In this paper, we will follow an approach for the combination of forecasts based on stochastic dominance (SD) analysis, and we test whether a simple average combination of forecasts would outperform forecast combinations with more elaborate weights. In this context, we will examine whether an equally weighted forecast combination is optimal when we analyze the forecast error distribution. Rather than assigning arbitrary equal weights to each forecast, we use stochastic dominance efficiency (SDE) analysis to propose a weighting scheme that dominates the equally weighted forecast combination or, alternatively, is dominated by it.

Typically in the literature SD comparisons are conducted in a pairwise manner. Barrett and Donald (2003) developed pairwise stochastic dominance comparisons that relied on Kolmogorov-Smirnov-type tests developed within a consistent testing environment. This offers a generalization of Beach and Davidson (1983), Anderson (1996), Davidson and Duclos (2000), who examined second-order SD using tests that rely on pairwise comparisons made at a fixed number of arbi-

[1] Smith and Wallis (2009) found that the finite sample error is the reason behind the forecast combination puzzle. Aiolfi et al. (2011) suggested that potential improvements can be made by using a simple equal-weighted average of forecasts from various time-series models and survey forecasts. See also Diebold and Pauly (1987), Clements and Hendry (1998, 1999, 2006), and Timmermann (2006) for a discussion of model instability and Elliot and Timmermann (2005) forecast combinations for time-varying data.

trarily chosen points, an undesirable feature that may lead to a test inconsistency. Linton et al. (2005) propose a subsampling method that can address both dependent samples and dependent observations within samples. This is appropriate for conducting SD analysis for model selection among many forecasts. In this context, comparisons are available for pairs for which one can compare one forecast with another forecast and conclude whether one forecast dominates the other. Hence, one can find the best individual model by comparing all forecasts. Lately, multivariate (multidimensional) comparisons have become more popular. In an application to optimal portfolio construction in finance, Scaillet and Topaloglou (2010), ST hereafter, used SD efficiency tests that can compare a given portfolio with an optimally diversified portfolio constructed from a set of assets.[2] We adapt the SDE methodology into a forecasting setting to obtain the optimal (worst) forecast combination. The main contribution of the paper is the derivation of an optimal (and a worst) forecast combination based on SDE analysis with differential forecast weights. For the optimal (worst) forecast combination, this forecast combination will minimize (maximize) the number of forecast errors that surpass a given threshold. In other words, we will examine the forecast error distribution of the average forecast combination at different parts of the empirical distribution and test whether the average forecast combination is optimal (worst) for all sections of the forecast error distribution. Furthermore, we investigate whether there is an alternative forecast combination that can offer an optimal (worst) forecast combination at some parts of the forecast error distribution.

To achieve SD, we maximize the difference between two cumulative distribution functions. This maximization results in the worst forecast combination constructed from the set of forecast models in the sense that it reaches the maximum value of forecast errors for a given probability level. In other words, at a given level of the forecast distribution, there will be always more forecast errors with the worst forecast combination when compared to the average forecast combination. Similarly, reversing the two cumulative forecast error distributions (the average and alternative one) would result in the optimal forecast combination, where the forecast combination now achieves the minimum value of forecast errors for a given probability level. We expand on this point in the next section.

This paper differs from the forecast combination literature insofar as most papers in the latter consider the minimization of the total sum of the squared forecast errors (or the mean squared forecast errors), something that relies only on the first two moments of the forecast error distribution. However, in this paper, we analyze the entire forecast error distribution, which takes into account all moments. Rather than relying on point estimates (or single optimal forecast combina-

[2]In a related paper, Pinar et al. (2013) used a similar approach to construct an optimal Human Development Index. The same methodology was applied in Agliardi et al. (2012), where an optimal country risk index was constructed following SD analysis with differential component weights, yielding an optimal hybrid index for economic, political, and financial risk indices that do not rely on arbitrary weights as rating institutions do.

tions), we derive the optimal and worst forecast combinations at different parts of the empirical forecast error distribution. In other words, rather than choosing the one forecast combination that minimizes the sum of the squared forecast errors, we derive different combinations that will minimize (maximize) the cdf of forecast errors up to a given forecast error level. For instance, a forecast combination might work very well for 95% of forecasts but perform relatively poorly for the remaining 5%. Methods that minimize the sum of the squared forecast errors might not choose this forecast combination as their optimal solution; however, if one were to allow for the information provided by higher moments, this combination would have been chosen. Hence, if the loss function or forecast error distribution is skewed, different forecast combinations might work better at different areas of the empirical distribution of the forecast errors (see, e.g., Elliott and Timmermann, 2004). The SDE methodology takes into account all moments of the empirical forecast error distribution; consequently, it provides the best (worst) forecast combination at different areas of the empirical distribution.

For our empirical applications, we use two exchange rate series given in a weekly frequency for the Japanese Yen/U.S. Dollar and U.S. Dollar/Great Britain Pound for the period from 1975 to 2010. We derive forecast combinations obtained with SDE at different parts of the forecast error distribution for the worst and optimal forecast combination for each application and forecast horizons. Overall, we find that the optimal forecast combinations with SDE weights perform better than different forecast selection and combination methods for the majority of the cases. However, there are also some very few cases where some other forecast selection and combination model performs equally well or even better at some part of the forecast error distribution. For example, at some percentiles of the forecast error distribution, the equally weighted forecast combination offers the best combination when compared with other possible combinations, something that provides partial support for the forecast combination puzzle.[3] We also find that the worst forecast combination with SDE weights is the worst one at all parts of the forecast error distribution when compared with other forecast selection and combination models. For the worst forecast combination with SDE weights, the random walk (RW hereafter) model is the model that consistently contributes with considerably more to the worst forecast combination for all variables being forecasted, for all forecast horizons and for the entire support of the forecast error empirical distribution. For the optimal forecast combination obtained with SDE weights, the best forecasting model (i.e., the model that gets relatively more weight than other forecasting models) includes different sets of models at different parts of the empirical distribution. On average, flexible neural network autoregressive (NNETTS) models and self-exciting threshold autoregressive (SETAR) models work well in most of the support of the empirical forecast error distribution.

The remainder of the paper includes the following. In section 2, we define the concept of SDE

[3]To observe full support for the puzzle, we would expect the equally weighted combination to perform best over all parts of the forecast error distribution. On the other hand, for our empirical applications to refute the puzzle, we would expect the simple average combination to be the worst.

and discuss the general hypothesis for SDE at any order. Section 3 describes the data, time-series forecasting models and forecast methods used in our paper as well as alternative forecast selection and combination methods. Section 4 presents the empirical analysis where we use the SDE methodology to find the optimal (worst) forecast combination for macroeconomic variables with different forecast horizons and compare these findings with those from the other forecast selection and combination methods. Finally, section 5 concludes the paper.

2 Hypothesis, Test Statistics and Asymptotic Properties

Let us start with data $\{y_t; t \in \mathbb{Z}\}$ and the $(m \times 1)$ column vector of forecasts $\left\{\widehat{\boldsymbol{y}}_{t+h,t}; t, h \in \mathbb{Z}\right\}$ for y_{t+h} obtained from m different forecasting models generated at time t for the period of $t+h$ $(h \geq 1)$, where h is the forecast horizon and T is the final forecasting period. Furthermore, let y_{t+h} denote the actual values over the same forecast period.

The equally weighted column vector, $\boldsymbol{\tau}$, is used to obtain the simple average of individual forecasts derived from the m different models, i.e., $\widehat{y}^{ew}_{t+h,t} = \boldsymbol{\tau}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$, where $\boldsymbol{\tau}$ is the $(m \times 1)$ column vector with entries $\frac{1}{m}$'s. The forecast error of the equally weighted forecast combination is given by $\hat{\varepsilon}^{ew}_{t+h,t} = y_{t+h} - \widehat{y}^{ew}_{t+h,t}$, which defines some loss function L, $L = L(\varepsilon_{t+h})$. Let us now consider an alternative weighting column vector $\boldsymbol{\lambda} \in \mathbb{L}$, where $\mathbb{L} := \{\boldsymbol{\lambda} \in \mathbb{R}^n_+ : \boldsymbol{e}^{'}\boldsymbol{\lambda} = 1\}$ with e being a vector of ones. With this alternative weighting scheme, one can obtain a forecast combination, i.e., $\widehat{y}^{w}_{t+h,t} = \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$. Similarly, this forecast combination with this alternative weighting scheme will generate a loss function, $\hat{\varepsilon}^{w}_{t+h,t} = y_{t+h} - \widehat{y}^{w}_{t+h,t}$.

Note that we can have different loss functions depending on the different choices of weights available. The forecast combination literature employs various objective functions derived from the loss function to obtain optimal weights, where the loss for a weighted forecast combination is given as $\hat{\varepsilon}^{w}_{t+h,t}$ (see, e.g., Hyndman and Koehler, 2006, for an extensive list of accuracy measures). It is common in the literature to use the norm of the forecasting error to find the optimal weights (see Timmermann, 2006).

In other words, the optimal forecast combination, $\boldsymbol{\lambda}^{*}_{t+h,t}$, is obtained by solving the problem

$$\boldsymbol{\lambda}^{*}_{t+h,t} = \arg\min_{\boldsymbol{\lambda}} E\left[L(\hat{\varepsilon}^{w}_{t+h,t}(\boldsymbol{\lambda}_{t+h,t}) \mid \widehat{\boldsymbol{y}}_{t+h,t}\right] \qquad s.t. \qquad \boldsymbol{e}^{'}\boldsymbol{\lambda} = 1 \tag{1}$$

where the expectation is taken over the conditional distribution of $\varepsilon_{t+h,t}$.

However, it is well know that all of the moments of the forecast error distribution will affect the combination of weights (see, e.g., Geweke and Amisano, 2011), and if one were to find the optimal weights by analyzing the entire distribution of the errors, this would lead to a more informative outcome. In this paper, SDE analysis allows for all moments to be considered as it examines the entire forecast error distribution. For example, if one were to find weights by minimizing the mean squared forecast errors (MSFE) and the forecast distribution was asymmetric with some important

outliers, then the weighted combination, which would have been obtained as the solution, would have ignored these important features of the empirical distribution. In other words, under an MSFE loss function, the optimal forecast combination is obtained by the optimal trade-off between squared bias and the forecast error variance (i.e., the optimal forecast combination only depends on the first two moments of the forecast errors). However, if the forecast error distribution is skewed, different weighted combinations would work better at different parts of the empirical distribution of the forecast errors (see, e.g., Elliott and Timmermann, 2004). Hence, looking at all of the moments of the forecast error would result in more robust weighting schemes.

For this paper, we follow a loss function that depends on the forecast error $L(\varepsilon_{t+h,t})$ that has the following properties (Granger, 1999):

$i. L(0) = 0,$

$ii. \min_{e} L(\varepsilon) = 0$, i.e., $L(\varepsilon) \geq 0$,

$iii. L(\varepsilon)$ is monotonic non-decreasing as ε moves away from 0:

i.e., $L(\varepsilon_1) \geq L(\varepsilon_2)$ if $\varepsilon_1 > \varepsilon_2 \geq 0$ and if $\varepsilon_1 < \varepsilon_2 \leq 0$.

(i) suggests that there is no loss when there is no error, (ii) suggests that the minimum loss is zero, and finally, (iii) suggests that the loss is determined by its distance to zero error irrespective of its sign.[4] This loss function may have further assumptions, such as being symmetric, homogenous, or differentiable up to some order (see Granger, 1999, for the details).

In this paper, we take $\hat{\varepsilon}^{ew}_{t+h,t}$ as the random variable of interest, and we test whether it is stochastically dominant vis-à-vis all other forecast combinations. In other words, we test whether the cumulative distribution function of the forecast errors of the equally weighted forecast combination is stochastically efficient. Let us define $F(\hat{\varepsilon}^{ew}_{t+h,t})$ as the continuous cdf of $\hat{\varepsilon}^{ew}_{t+h,t} = (y_{t+h} - \boldsymbol{\tau}^{'}\widehat{\boldsymbol{y}}_{t+h,t})$, where $\hat{\varepsilon}^{ew}_{t+h,t}$ is the random variable of forecast errors that is obtained by combining forecasting models with weights $\boldsymbol{\tau}$, being $\frac{1}{m}$s. On the other hand, if one were to consider an alternative weighting vector, $\boldsymbol{\lambda}$, i.e., all the forecasts that are obtained from different models have non-negative weights. We denote by $F(\hat{\varepsilon}^{w}_{t+h,t})$ the continuous cdf of $\hat{\varepsilon}^{w}_{t+h,t} = (y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t})$. Furthermore, let us denote by $G(z, \boldsymbol{\tau}; F)$ and $G(z, \boldsymbol{\lambda}; F)$ the cdfs of the forecast errors from the forecast combination of $\boldsymbol{\tau}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ and $\boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ at point z given $G(z, \boldsymbol{\tau}; F) := \int_{\mathbb{R}^n} \mathbb{I}\{(y_{t+h} - \boldsymbol{\tau}^{'}\widehat{\boldsymbol{y}}_{t+h,t}) \leq z\} dF(\hat{\varepsilon}^{ew}_{t+h,t})$ and $G(z, \boldsymbol{\lambda}; F) := \int_{\mathbb{R}^n} \mathbb{I}\{(y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}) \leq z\} dF(\hat{\varepsilon}^{w}_{t+h,t})$, Respectively, where z represents the forecast error level[5] and $\mathbb{I}$ represents the indicator function (Davidson and Duclos 2000).

For any two forecast error distributions, we say that the forecast combination $\boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ dominates the distribution of the equally weighted forecast combination $\boldsymbol{\tau}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ stochastically at first order (SD1) if, for any point z of the distribution, $G(z, \boldsymbol{\tau}; F) \geq G(z, \boldsymbol{\lambda}; F)$. In general, the dominant combination refers to a "best outcome" case because there is more mass to the right of z with

[4] In this paper, we take the absolute values of negative errors when we concentrate on the magnitude of the distance from zero error, not the sign of errors.

[5] As suggested by the assumptions above, we concentrate on the magnitude of the errors; and therefore, z represents the monotonic non-decreasing distance to zero error.

$G(z, \boldsymbol{\lambda}; F)$ than with $G(z, \boldsymbol{\tau}; F)$. In the context of the present analysis, because the distribution of outcomes refers to forecast errors, the "best outcome" dominant case corresponds to a forecast combination with the largest forecast errors up to an error level z and, as such, it would yield the worst possible forecast combination. More precisely, in the context of our analysis, if z denotes the forecast error level, then the inequality in the definition means that the value (mass) of the cdf of the forecast errors obtained with the forecast combination of $\boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ at point z is no larger than the value (mass) of the cdf of the forecast errors with the equally weighted forecast combination, $\boldsymbol{\tau}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$. In other words, there is at least as high a proportion of forecast error with the forecast combination of $\boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ as with the equally weighted forecast combination, $\boldsymbol{\tau}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$. If the forecast combination $\boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ dominates the equally weighted forecast combination $\boldsymbol{\tau}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ at the first order, then $\boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ yields the worst forecast combination for that given forecast error level, z.

More precisely, to achieve stochastic dominance, we maximize the following objective function: Max $[G(z, \boldsymbol{\tau}; F) - G(z, \boldsymbol{\lambda}; F)]$. This maximization results in the worst forecast combination, $\boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$, that can be constructed from the set of forecast models in the sense that it reaches the maximum value of forecast errors for a given probability level (i.e., the highest number of forecast errors that are above a given z level). Following the traditional definition of SD, we center our discussion around the worst forecast combination, consistent with the highest mass of forecast errors. Any dominant combination in that case would imply a worse forecast combination than the one that it dominates. If there is no other forecast combination with an alternative weighting scheme that dominates the equally weighted forecast combination, then the latter would constitute the worst forecast combination for that given forecast error level. Of course, if one were to reverse the two cumulative forecast error distributions that are given in the maximization problem above, i.e., Max $[G(z, \boldsymbol{\lambda}; F) - G(z, \boldsymbol{\tau}; F)]$, such a reversal would result in the optimal (best) forecast combination. In that case, an alternative forecast combination would offer the fewest number of forecast errors that are above a given forecast error, z, and would therefore be consistent with minimal forecast errors. In the empirical section, we will derive worst- and best-case forecast combinations, and we will assess whether the equally weighted forecast combination is the best (or worst) forest combination for different sections of the empirical distribution.

The general hypotheses for testing whether the equally weighted forecast combination, $\boldsymbol{\tau}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$, is the worst forecast combination at the stochastic dominance efficiency order of j, hereafter SDE_j, can be written compactly as:

$$
\begin{aligned}
H_0^j &: \mathcal{J}_j(z, \boldsymbol{\tau}; F) \leq \mathcal{J}_j(z, \boldsymbol{\lambda}; F) \quad \text{for all } z \in \mathbb{R} \text{ and for all } \boldsymbol{\tau},\ \boldsymbol{\lambda} \in \mathbb{L}, \\
H_1^j &: \mathcal{J}_j(z, \boldsymbol{\tau}; F) > \mathcal{J}_j(z, \boldsymbol{\lambda}; F) \quad \text{for some } z \in \mathbb{R} \text{ or for some } \boldsymbol{\tau},\ \boldsymbol{\lambda} \in \mathbb{L}.
\end{aligned}
$$

where

$$\mathcal{J}_j(z, \boldsymbol{\lambda}; F) = \int_{\mathbb{R}^n} \frac{1}{(j-1)!}(z - (y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}))^{j-1}\mathbb{I}\{(y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}) \leq z\}dF(\varepsilon^{w}_{t+h,t}) \quad (2)$$

or, equivalently,

$$\mathcal{J}_j(z, \boldsymbol{\lambda}; F) = \int_{\mathbb{R}^n} \frac{1}{(j-1)!}(z - \varepsilon^{w}_{t+h,t})^{j-1}\mathbb{I}\{\varepsilon^{w}_{t+h,t} \leq z\}dF(\varepsilon^{w}_{t+h,t}) \quad (3)$$

and $\mathcal{J}_1(z, \boldsymbol{\lambda}; F) := G(z, \boldsymbol{\lambda}; F)$.[6] Under the null hypothesis H_0^j there is no distribution of forecast errors obtained from any alternative forecast combination $\boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ that dominates the distribution. Under the alternative hypothesis H_1^j, we can construct a forecast combination $\boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ for which, for some forecast error level z, the function $\mathcal{J}_j(z, \boldsymbol{\tau}; F)$ is greater than the function $\mathcal{J}_j(z, \boldsymbol{\lambda}; F)$. Thus, $j = 1$, the equally weighted forecast combination $\boldsymbol{\tau}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ is stochastically dominated (i.e., does not yield the worst forecast combination) at the first order if and only if some other forecast combination $\boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ dominates it at some forecast error level z. In other words, there is an alternative weighting scheme, $\boldsymbol{\lambda}$, such that when forecasts are combined with these weights, $\boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ yields a distribution of forecast errors that are higher than that of weighted forecast combination. Alternatively, the equally weighted forecast combination $\boldsymbol{\tau}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ is stochastically dominant (i.e., is the worst forecast combination) at the first order if and only if there is no other forecast combination $\boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ with an alternative weighting scheme that dominates it at all forecast error levels. In other words, at any forecast error level, the equally weighted forecast combination always yields a distribution of forecast errors such that it always gives a higher probability of more forecast errors.

We obtain SD at the first and second orders when $j = 1$ and $j = 2$, respectively. The hypothesis for testing the SD of order j of the distribution of the equally weighted forecast combination $\boldsymbol{\tau}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ over the distribution of an alternative forecast combination $\boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ takes analogous forms but uses a single given $\boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ rather than several of them.

The empirical counterpart of (3) is simply obtained by integrating with respect to the empirical distribution $\hat{F}$ of F, which yields the following:

$$\mathcal{J}_j(z, \boldsymbol{\lambda}; \hat{F}) = \frac{1}{N_f}\sum_{N_f=1}^{N_f} \frac{1}{(j-1)!}(z - (y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}))^{j-1}\mathbb{I}\{(y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}) \leq z\} \quad (4)$$

[6] Defining

$$\mathcal{J}_2(z, \boldsymbol{\lambda}; F) := \int_{-\infty}^{z} G(u, \boldsymbol{\lambda}; F)du = \int_{-\infty}^{z} \mathcal{J}_1(u, \boldsymbol{\lambda}; F)du,$$
$$\mathcal{J}_3(z, \boldsymbol{\lambda}; F) := \int_{-\infty}^{z}\int_{-\infty}^{u} G(v, \boldsymbol{\lambda}; F)dvdu = \int_{-\infty}^{z} \mathcal{J}_2(u, \boldsymbol{\lambda}; F)du,$$

and so on. Davidson and Duclos (2000) (Equation (2)) have stated that

$$\mathcal{J}_j(z, \boldsymbol{\lambda}; F) = \int_{-\infty}^{z} \frac{1}{(j-1)!}(z-u)^{j-1}dG(u, \boldsymbol{\lambda}, F),$$

which can be rewritten as it appears in the text.

where N_f is the number of factor of realizations.[7] We consider the weighted Kolmogorov-Smirnov type test statistic

$$\hat{S}_j := \sqrt{N_f}\frac{1}{N_f}\sup_{z,\boldsymbol{\lambda}}\left[\mathcal{J}_j(z,\boldsymbol{\tau};\hat{F}) - \mathcal{J}_j(z,\boldsymbol{\lambda};\hat{F})\right] \tag{5}$$

and a test based on the decision rule

$$\text{“ Reject } \quad H_0^j \quad \text{ if } \quad \hat{S}_j > c_j \text{ ”},$$

where c_j is some critical value (see ST for the derivation of the test statistic).

To make the result operational, we need to find an appropriate critical value c_j. Because the distribution of the test statistic depends on the underlying distribution, this is not an easy task, and we decide hereafter to rely on a block bootstrap method to simulate p-values, where the critical values are obtained using a supremum statistic (see Appendix).[8] The test statistic $\hat{S}_1$ for first-order stochastic dominance efficiency is derived using mixed integer programming formulations (see Appendix).[9]

3 Empirical Analysis

3.1 Data, Forecasting Models, and Forecast Methodology

In this section, we apply the SDE testing methodology to obtain optimal (worst) forecast combinations on Japanese yen/U.S. dollar and U.S. dollar/Great Britain pound exchange rate returns data. We use log first differences of the exchange rate levels. The exchange rate series data are expressed with a weekly frequency for the period between 1975:1-2010:52.[10] The use of weekly data avoids the so-called weekend effect, as well as other biases associated with non-trading, bid-ask spread, asynchronous rates and so on, which are often present in higher-frequency data. To initialize our parameter estimates, we use weekly data between 1975:1 - 2006:52. We then generate

[7]Forecasts from different models are updated recursively by expanding the estimation window by one observation forward, thereby reducing the pseudo-out-of-sample test window by one period. Therefore, for each of h-step forecasts, we calculate N_f forecast errors for each of the models, as explained in the following section.

[8]The asymptotic distribution of $\hat{F}$ is given by $\sqrt{N_f}(\hat{F}-F)$, which tends weakly to a mean zero Gaussian process $B\circ F$ in the space of continuous functions on R^n (see, e.g., the multivariate functional central limit theorem for stationary strongly mixing sequences stated in Rio (2000)). ST (2010) derive the limiting behavior by using the Continuous Mapping Theorem (as in Lemma 1 of Barrett and Donald (2003)); see ST (2010) Lemma 2.1.

[9]In this paper, we only test first-order SD in the empirical applications below. Because there are forecast combinations with alternative weighting schemes that dominate (or are dominated by) the equally weighted forecast combination at the first order, we do not move to the second one.

[10]The daily noon buying rates in New York City certified by the Federal Reserve Bank of New York for customs and cable transfer purposes are obtained from the FREDÂ® Economic Data system of Federal Reserve Bank of St. Louis (http://research.stlouisfed.org). The weekly series is generated by selecting the Wednesday series (if Wednesday is a holiday, then the subsequent Thursday is used).

pseudo- out-of-sample forecasts of 2007:1 - 2010:52. Parameter estimates are updated recursively by expanding the estimation window by one observation forward and thereby reduce the pseudo out-of-sample test window by one period.

In our out-of-sample forecasting exercise, we concentrate exclusively on univariate models, and we consider three types of linear univariate models and four types of nonlinear univariate models. The linear models are random walk (RW), autoregressive (AR), and autoregressive moving-average (ARMA) models; the nonlinear ones are logistic smooth transition autoregressive (LSTAR), self-exciting threshold autoregressive (SETAR), Markov-switching autoregressive (MS-AR), and autoregressive neural network (NNETTS) models.

Let $\hat{y}_{t+h,t}$ be the forecast of y_t that is generated at time t for the time $t+h$ ($h \geq 1$) by any forecasting model. In the RW model, $\hat{y}_{t+h,t}$ is equal to the value of y_t at time t.

The ARMA model is

$$y_t = \alpha + \sum_{i=1}^{p} \phi_{1,i} y_{t-i} + \sum_{i=1}^{q} \phi_{2,i} \varepsilon_{t-i} + \varepsilon_t, \tag{6}$$

where p and q are selected to minimize the Akaike Information Criterion (AIC) with a maximum lag of 24. After estimating the parameters of equation (6), one can easily produce h-step ($h \geq 1$) forecasts through the following recursive equation:

$$\hat{y}_{t+h,t} = \alpha + \sum_{i=1}^{p} \hat{\phi}_{1,i} y_{t+h-i} + \sum_{i=1}^{q} \hat{\phi}_{2,i} \varepsilon_{t+h-i}\,. \tag{7}$$

When $h > 1$, to obtain forecasts, we iterate a one-period forecasting model by feeding the previous period forecasts as regressors into the model. This means that when $h > p$ and $h > q$,, y_{t+h-i} is replaced by $\hat{y}_{t+h-i,t}$ and ε_{t+h-i} by $\hat{\varepsilon}_{t+h-i,t} = 0$.

An obvious alternative to iterating forward on a single-period model would be to tailor the forecasting model directly to the forecast horizon, i.e., to estimate the following equation by using the data up to t:

$$y_t = \alpha + \sum_{i=0}^{p} \phi_{1,i} y_{t-i-h} + \sum_{i=0}^{q} \phi_{2,i} \varepsilon_{t-i-h} + \varepsilon_t, \tag{8}$$

for $h \geq 1$. We use the fitted values of this regression to directly produce an h-step ahead forecast.[11]

Because it is a special case of ARMA, the estimation and forecasts of the AR model can be obtained by simply setting $q = 0$ in (6) and (8).

[11] Deciding whether the direct or the iterated approach is better is an empirical matter because it involves a trade-off between the estimation efficiency and the robustness-to-model misspecification; see Elliot Timmermann (2008). Marcellino et al. (2006) have addressed these points empirically using a dataset of 170 US monthly macroeconomic time series. They have found that the iterated approach generates the lowest MSE values, particularly if lengthy lags of the variables are included in the forecasting models and if the forecast horizon is long.

The LSTAR model is

$$y_t = \left(\alpha_1 + \sum_{i=1}^{p} \phi_{1,i}\, y_{t-i}\right) + d_t \left(\alpha_2 + \sum_{i=1}^{q} \phi_{2,i}\, y_{t-i}\right) + \varepsilon_t, \tag{9}$$

where $d_t = (1 + \exp\{-\gamma(y_{t-1} - c)\})^{-1}$. Whereas ε_t are regarded as normally distributed i.i.d. variables with zero mean, α_1, α_2, $\phi_{1,i}$, $\phi_{2,i}$, γ and c are simultaneously estimated by maximum likelihood methods.

In the LSTAR model, the direct forecast can be obtained in the same manner as with ARMA, which is also the case for all of the subsequent nonlinear models[12], but it is not possible to apply any iterative scheme to obtain forecasts for multiple steps in advance, as can be done in the case of linear models. This impossibility follows from the general fact that the conditional expectation of a nonlinear function is not necessarily equal to a function of that conditional expectation. In addition, one cannot iteratively derive the forecasts for the time steps $h > 1$ by plugging in the previous forecasts (see, e.g., Kock and Terasvirta, 2011).[13] Therefore, we use the Monte Carlo integration scheme suggested by Lin and Granger (1994) to numerically calculate the conditional expectations, and we then produce the forecasts iteratively.

When $|\gamma| \to \infty$, the LSTAR model approaches the two-regime SETAR model, which is also included in our forecasting models. As with LSTAR and most nonlinear models forecasting with SETAR does not permit the use a simple iterative scheme to generate multiple-period forecasts. In this case, we employ a version of the Normal Forecasting Error (NFE) method suggested by Al-Qassam and Lane (1989) to generate multistep forecasts.[14] NFE is an explicit, form-recursive approximation for calculating higher-step forecasts under the normality assumption of error terms and has been shown by De Gooijer and De Bruin (1998) to perform with reasonable accuracy compared with numerical integration and Monte Carlo method alternatives.

The two-regime MS-AR model that we consider here is as follows:

$$y_t = \alpha_s + \sum_{i=1}^{p} \phi_{s,i} y_{t-i} + \varepsilon_t, \tag{10}$$

where s_t is a two-state discrete Markov chain with $S = \{1, 2\}$ and $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$. We estimate MS-AR using the maximum likelihood expectation-maximization algorithm.

Although MS-AR models may encompass complex dynamics, point forecasting is less complicated in comparison to other non-linear models. The h-step forecast from the MS-AR model is

[12]This process involves replacing y_t with y_{t+h} on the left-hand side of equation (4) and running the regression using data up to time t to fitted values for corresponding forecasts.

[13]Indeed, d_t is convex in y_{t-1} whenever $y_{t-1} < c$, and $-d_t$ is convex whenever $y_{t-1} > c$. Therefore, by Jensen's inequality, naive estimation underestimates d_t if $y_{t-1} < c$ and overestimates d_t if $y_{t-1} > c$.

[14]A detailed exposition of approaches for forecasting from a SETAR model can be found in van Dijk at al. (2003)

$$\begin{aligned}\hat{y}_{t+h,t} &= P\left(s_{t+h}=1 \mid y_t,...,y_0\right)\left(\alpha_{s=1}+\sum_{i=1}^{p}\hat{\phi}_{s=1,i}y_{t+h-i}\right)\\ &\quad +P\left(s_{t+h}=2 \mid y_t,...,y_0\right)\left(\alpha_{s=2}+\sum_{i=1}^{p}\hat{\phi}_{s=2,i}y_{t+h-i}\right),\end{aligned} \tag{11}$$

where $P\left(s_{t+h}=i \mid y_t,...,y_0\right)$ is the ith element of the column vector $\mathbf{P}^h\hat{\xi}_{t|t}$. In addition, $\hat{\xi}_{t|t}$ represents the filtered probabilities vector and $\mathbf{P}^h$ is the constant transition probability matrix (see, Hamilton (1994)). Hence, multistep forecasts can be obtained iteratively by plugging in $1,2,3,\ldots$-period forecasts that are similar to the iterative forecasting method of the AR processes.

ARNN, which is the autoregressive single-hidden-layer feed-forward neural network model[15] suggested in Terasvirta (2006), is defined as follows:

$$y_t=\alpha+\sum_{i=1}^{p}\phi_i y_{t-i}+\sum_{j=1}^{h}\lambda_j d\left(\sum_{i=1}^{p}\gamma_i y_{t-i}-c\right)+\varepsilon_t, \tag{12}$$

where d is the logistic function, which is defined above as $d=\left(1+\exp\{-x\}\right)^{-1}$. In general, the estimation of an ARNN model may be computationally challenging. Here, we follow the QuickNet method, which is a type of "relaxed greedy algorithm"; it was originally suggested by White (2006). In contrast, the forecasting procedure for NNETTS is identical to the procedure for LSTAR.

To obtain pseudo-out-of-sample forecasts for a given horizon h, the models are estimated by running regressions with data that were collected no later than the date $t_0<T$, where t_0 refers to the date when the estimation is initialized, and T refers to the final date in our data. The first h-horizon forecast is obtained using the coefficient estimates from the initial regression. Next, after moving forward by one period, the procedure is repeated. For each h-step forecast, we calculate N_f $(=T-t_0-h-1)$ forecast errors for each of the models that we use in our applications.

3.2 Forecast selection and combination

Before proceeding with our application, in this section we offer different set of model selection and combination methods that are employed extensively in the literature. Akaike's information criterion (AIC) and Bayesian information criterion (BIC) are two of the most commonly used selection criteria that serve to select a forecasting model (see, for example, Swanson and Zeng, 2001; Drechsel and Maurin, 2010, among many others). The model that provides the lowest AIC or BIC, calculated as below, for a model m is chosen as the preferred model.

$$AIC(m)=n\ln(\widehat{\sigma}_m^2)+2k_m, \tag{13}$$

[15] see Franses and Dijk (2000) for a review of feed-forward-type neural network models.

$$BIC(m) = n\ln(\widehat{\sigma}_m^2) + k_m \ln n, \tag{14}$$

where $\widehat{\sigma}_m^2$ is the forecast error variance estimate and k_m is the number of regressors used in each respective model. This procedure requires the selection of the forecasting model that offers the minimum value of AIC or BIC. Another classical method that is used to select the best individual forecasting model is to select the model that offers the least forecast variance, also called predictive least squares (PLS) (Rissanen, 1986).

However, these procedures neglect the fact that, as is discussed above, the combination of different models could perform better than the selection of a single model as the best model. Therefore, the procedure can be modified accordingly so that weights given to each model is determined based on the distance between each model's AIC (BIC) from the minimal performing model's AIC (BIC) level. Hence, defining the difference between the $AIC(m)$ ($BIC(m)$) and the $min(AIC)$ (i.e., the model that offers the lowest AIC) as $\Delta AIC(m) = AIC(m) - \min(AIC)$ ($\Delta BIC(m) = BIC(m) - \min(BIC)$), the exponential "Akaike weights", $\boldsymbol{w}_{AIC}(m)$, (see, e.g., Burnham and Anderson, 2002) and "Bayesian weights", $\boldsymbol{w}_{BIC}(m)$, (see, e.g., Raftery, 1995; Fernández et al., 2001; Sala-i-Martin et al., 2004 among many others) can be obtained as follows:

$$\boldsymbol{w}_{AIC}(m) = \frac{\exp\left(-\frac{1}{2}\Delta AIC(m)\right)}{\sum_{j=1}^{M}\exp\left(-\frac{1}{2}\Delta AIC(j)\right)}, \tag{15}$$

$$\boldsymbol{w}_{BIC}(m) = \frac{\exp\left(-\frac{1}{2}\Delta BIC(m)\right)}{\sum_{j=1}^{M}\exp\left(-\frac{1}{2}\Delta BIC(j)\right)}, \tag{16}$$

Then these weights can be utilized to combine the forecasts of m models. Another commonly used method to combine forecasts is to allocate weights to each model inversely proportional to the estimated forecast error variances (Bates and Granger,1969).

Among all these model selection and combination methods, the recent literature, as mentioned earlier, also employs the equally weighted forecast combination and the median forecast (see e.g., Stock and Watson, 2004; Kolassa, 2011). All forecast model selection and combination methods discussed in this section will be employed and compared to the method with SDE weights proposed in this paper.

4 Results for the efficiency of forecast combinations

This section presents our findings of the tests for first-order SD efficiency of the equally weighted forecast combination. We find that the equally weighted forecast combination is not the worst forecast combination at any part of the forecast error distribution, whereas it is the best forecast combination at some parts of the distribution. We obtain the best and worst forecast combinations

of the model-based forecasts for the Japanese yen/U.S. dollar and the U.S. dollar/Great Britain pound exchange rate forecasts by computing the weighting scheme on each forecast model that offers the optimal (worst) forecast combination at different parts of the forecast error distribution.

In our applications, because the forecast error distribution with the equally weighted forecast combination is known, we can obtain the number of forecast errors that are less than each given level of forecast error, z. We test the full support of the empirical forecast error distribution of the average forecast combination, that is, we test whether the forecast error levels obtained with the equally weighted forecast combination is the worst (or best) forecast combination in its entirety against the alternative that there is an alternative weighting scheme that will offer the worst (or best) forecast combination at different parts of the empirical distribution. In the next section, we report the worst and optimal forecast combination for different percentiles (i.e., 50th, 75th, 95th percentiles) of the forecast error distribution for the two applications and for different horizons. We also report the average of the optimal and worst forecast combinations that are obtained for different forecast error levels.[16] For each application, we also compare the worst and best forecast combinations obtained with SDE weights with different set of model selection and combinations that are used commonly in the literature.

4.1 The Japanese yen/U.S. dollar exchange rate application

First, we begin our empirical analysis with the weekly Japanese yen/U.S. dollar exchange rate forecasts for different forecast horizons. We proceed with testing whether the equally weighted forecast combination of the forecasting models for different horizons is the worst (optimal) forecast combination at all forecast error levels or there are alternative weights on the forecast models that stochastically dominate (are dominated by) the equally weighted forecast combination, $\boldsymbol{\tau}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$, in the first-order sense for some or all forecast error levels, where the number of forecast errors above a given forecast error level z is maximized (minimized).

Table 1 presents the results for the 50th, 75th, and 95th percentiles of the forecast error distribution of the equally weighted forecast combination for the different forecast horizons (h). The third column reports the forecast error levels of average (or equally weighted) forecast combinations at these particular percentiles. The following columns provide the weights of the underlying forecasting models for the worst- and best-case (optimal) forecast combinations at the 50th, 75th, and 95th percentiles of the equally weighted forecast error distribution.

In one step ahead forecast horizon, i.e., when $h = 1$, we have 208 forecasts for each of the different time-series models. As indicated in the first panel of Table 1, there is always an alterna-

[16]The empirical distribution of forecast errors consists of different forecast error levels, possibly exceeding 200 depending on the nature of the application. Therefore, rather than reporting the optimal and worst forecast combination for all forecast error levels, we only report selected percentiles and the averages of the optimal and worst forecast combinations that are obtained for the different levels of forecast error in the next sections. However, the full set of optimal and worst forecast combinations can be obtained upon request from authors.

tive forecast combination that generates the worst or optimal forecast combination for the 50th, 75th and 95th percentiles of the forecast error distribution. For example, at the 50th percentile of the forecast error distribution, when forecasts from NNETTS and RW obtain weights of 4.37% and 95.63%, respectively, this combination offers the worst forecast combination for this part of the distribution. When forecasts from NNETTS and SETAR are combined at 4.51% and 95.49%, respectively, this combination offers the best-case forecast combination for the 50th percentile of the forecast error distribution. Similarly, for the 75th percentile of the forecast error distribution, when forecasts from ARMA and RW obtain weights of 0.35% and 99.65%, respectively, this combination offers the worst forecast combination up to this percentile, and when forecasts from NNETTS and SETAR are combined with weights of 84.08% and 15.92%, respectively, this combination is the optimal one. Finally, the RW model alone produces the worst forecasts for the 95th percentile of the distribution, whereas when forecasts from NNETTS, RW and SETAR are combined with weights of 11.53%, 3.36%, and 85.11%, respectively, this produces the best forecast combination.

Overall, when $h = 1$, different forecast combinations generate the worst and best forecast combinations for different sections of the forecast error distribution. In general, forecasts from the RW model most contribute to the worst forecast combination for all sections of the forecast error distribution. On the other hand, forecasts from NNETTS and SETAR contribute most to the optimal forecast combination with different weights for different sections of the forecast error distribution.

We carried out the same application when we extended the forecast horizon for 6 months (26 weeks), 6 months (26 weeks) and a year (52 weeks) (i.e., $h = 27$, 53 and 105, respectively), where for each case, each model produces 182, 156, and 104 forecasts, respectively. For $h = 27$, we find similar results to those of $h = 1$ for the worst forecast combination such that the RW model most contributes to the worst forecast combination at the 50th, 75th and 95th percentiles (see the second row of Table 1). For the optimal forecast combination, at the 50th and 75th percentiles, the SETAR and NNETTS models most contribute to the optimal forecast combination. The equally weighted forecast combination is the optimal forecast combination at the 95th percentile of the forecast error distribution (see the second row of Table 1). The similar trend for the worst forecast combination continues for $h = 53$, but the LSTAR and RW models contribute relatively more at different areas of the forecast error distribution for the best forecast combination relative to the earlier cases. Finally, when $h = 105$, if forecasts from the NNETTS and RW models are combined with weights of 5.32% and 94.68%, respectively, this forecast combination gives the worst forecast combination for the 50th percentile of the forecast error distribution (see the fourth row of Table 1). However, if one were to combine forecast from the NNETTS and RW models with weights of 82.54% and 17.46%, respectively, this time, this forecast combination offers the optimal forecast combination. A similar result is obtained at the 75th percentile of the forecast

error distribution. In other words, two models with different weights can be combined for both the worst and best forecast combinations. Finally, at the 95th percentile of the forecast error distribution, the forecasts from the RW model and the equally weighted combination give the worst and best forecast combinations, respectively.

In this subsection, we presented the best and worst forecast combinations at different percentiles of forecast error distribution when we consider the equally weighted forecast combination as the "benchmark". In the next subsection, we offer a comparison of SDE weights not only with equally weighted forecast combination but also with median forecast, model selection methods (i.e., AIC, BIC, and PLS), and the forecast combination methods (i.e., combination of forecasts with Bates and Granger, AIC, and BIC weights).

4.2 Comparisons

SDE weights obtained in the previous section suggested that when the equally weighted forecast combination is the benchmark, there is always an alternative forecast combination which would constitute a worse cases at all parts of the error distribution for all forecast horizons. On the other hand, SDE method offers weighting schemes at all forecast horizons that offer better forecast combinations at 50th and 75th percentile of forecast error distribution when compared to the equally weighted forecast combination. The equally weighted forecast combination performs equally well as SDE weights at the 95th percentile of the error distribution for some forecast horizons. To evaluate SDE weights further, we also obtain median forecast, and forecasts with different model selection and combination methods that are mentioned above.

To make the results more apparent for each forecast horizon, Table 2 presents the number of forecast errors that are equal to or less than a given benchmark forecast error level, z, at the 50th, 75th and 95th percentiles with eqaully weighted equally weighted forecast combination (EW), median forecast (Median), best model chosen with AIC, BIC and PLS, and combination of models with Bates and Granger, AIC, and BIC weights.

In Table 2, we calculate the number of observations that are equal to or less than forecast error levels at 50th, 75th, and 95th percentile of the forecast error distribution from the equally weighted forecast combination. The worst and optimal forecast combination with the SDE weights are obtained using the weights from Table 1. Moreover, we obtain median forecast, forecasts from the model that is chosen with AIC, BIC, PLS, and forecast combinations with Bates and Granger, AIC, and BIC weights. Each of these methods yields forecast errors which are compared with the ones obtained with SDE weights at different sections of the forecast error distribution. For example, for $h = 1$, at 50th percentile of forecast error distribution, there are 104 forecast errors that are less than or equal to the level of forecast error of 0.0104 when forecasts are combined with equal weights. On the other hand, the best forecast combination with SDE weights yields 107 forecast errors that are equal to or less than 0.0104. Similarly, for the 75th and 95th percentiles,

the best forecast combination with SDE weights performs better than the other forecast selection methods as there are 159 and 200 forecasts that produce error levels that are equal to or less than 0.0172 and 0.0318, respectively. In other words, the optimal forecast combinations with SDE weights produce 49 and 8 forecasts that give forecast error levels that are above 0.0172 and 0.0318, respectively. Finally, the worst forecast combination with SDE weights always produces more forecast errors above a given forecast error when it is compared with the other forecast selection and combination methods. At the 50th percentile of the forecast error distribution, the equally weighted forecast combination is the second-best one, whereas the median forecast is the second-best at 75th percentile of the forecast error distribution. Overall, what we observe is that weights obtained with SDE method perform better when compared to others at different parts of the forecast error distribution.

We carry out the same analysis when we change the forecast horizons. When $h = 27$, at the 95th percentile of the forecast error distribution, using the forecast combinations with Bates and Granger, SDE weights and equally weighted forecast combination outperform the other methods. For $h = 105$, at the 95th percentile of the forecast error, all forecast selection and combination models perform equally well. The forecast combination with SDE weights perform better than the other forecast selection and combination models for all remaining cases.

We only presented the SDE weights for the best and worst forecast combination at 50th, 75th, and 95th percentiles of the forecast error distribution. However, Table 3 illustrates the average contribution of each forecasting model to the worst and best forecast combination with SDE weights. These average contributions are calculated by averaging the different weights over all percentiles of the entire distribution. One can see that each model contributes slightly to the worst forecast combination in different areas of the forecast error distribution for different forecast horizons; however, the main contributor to the worst forecast combination is the RW model. On the other hand, all models contribute to the optimal forecast combination at different areas of the distribution, but the NNETTS, SETAR and MS models are those that contribute more than others, with their average contributions being 57.07%, 24.68%, and 11.95%, respectively.

Overall, for the application to weekly Japanese yen/U.S. dollar exchange rate forecasts, we find that the worst forecast combinations with SDE weights never outperform any other forecast selection and combination models. On the other hand, best forecast combination with SDE weights mostly outperforms the other forecast selection and combination models, with some few exceptions. For example, the forecast combination with equal weights and that with the Bates and Granger weights perform equally well at the 95th percentile of the forecast error distribution. Overall, using SDE weights, the RW model contributes most to the worst forecast combination on average at the different areas of the forecast error distribution for different horizons, whereas the NNETTS and SETAR models have larger weights for the optimal forecast combination for almost all forecast horizons we considered.

4.3 U.S. dollar/Great Britain pound exchange rate application

In this subsection, we obtain the optimal (worst) forecast combination for the foreign exchange rate of U.S. dollar/Great Britain pound forecasts for different time horizons. Table 4 presents the worst and best forecast combinations with SDE method at the 50th, 75th and 95th percentiles of the forecast error distribution of the equally weighted forecast combination when $h = 1$, 27, 53 and 105, respectively. Table 5 reports the number of forecast errors below a given forecast error level with different forecast selection and combination methods for different percentiles of the forecast error distribution. Finally, Table 6 presents the average SDE weights of each model that contribute to the worst and optimal forecast combination.

The results obtained for the foreign exchange rate of U.S. dollar/Great Britain pound are very similar to the ones obtained for the Japanese yen/U.S. dollar exchange rate data. Table 5 summarizes the comparisons of performance of different models at different sections of the forecast error distribution for different horizons. We find that the worst forecast combinations with SDE weights never outperform any of other forecast selection and combination models at any part of the forecast error distribution. On the other hand, SDE weights for the best forecast combination outperforms the other models for $h = 1$ and $h = 105$ at 50th, 75th and 95th percentiles of the forecast error distribution. However, when $h = 27$, the best forecast model chosen with the predictive least squares performs equally well as the model with SDE weights at 50th percentile of the forecast error distribution. For $h = 53$, the best forecast combination with SDE weights performs equally well with some other forecast selection and combination models at 75th and 95th percentiles of the forecast error distribution, but the forecast combination with Bates and Granger weights performs the best at 50th percentile of the forecast error distribution. Overall, the optimal forecast combination with SDE weights performs better than its competitors except in very few cases.

The forecasts from the RW model contribute most to the worst forecast combination with SDE weights at all parts of the forecast error distribution, whereas, on average, forecasts from the NNETTS and SETAR models contribute most to the optimal forecast combination obtained with SDE (see Table 6). However, these models contribute differently at different parts of the forecast error distribution. For example, the SETAR model contributes the most to the optimal forecast combination when the forecaster gives importance to the 50th percentile of the distribution when $h = 1$, 27, and 53, whereas forecasts from the NNETTS model contributes more to the optimal forecast combination at higher percentiles of the forecast error distribution. Forecasts from the NNETTS and SETAR models consistently contribute to the best forecast combinations, whereas forecasts from the ARMA and MS models occasionally contribute with higher weights to the optimal forecast combinations at different parts of the forecast error distribution when different forecast horizons are used.

Overall, for the application to the weekly U.S. dollar/Great Britain pound exchange rate fore-

casts, the findings are very similar to those of Japanese yen/U.S. dollar exchange rate application. We find that the SDE weights for the worst forecast combination never outperforms any other model selection and combination for any part of the forecast error distribution. On the other hand, the best forecast combination with SDE weight overall performs better than other forecast selection and combination cases, except in a very few cases. Finally, we find that the NNETTS and SETAR models contribute with proportionately greater weights to the optimal forecast combination for different sections of the forecast error distribution for all forecast horizons. On the other hand, the RW model always contributes with the highest weight to the worst forecast combination with SDE weights at all parts of the forecast error distribution.

5 Conclusion

This paper studies the SDE properties to combine forecasts by which worst and optimal forecast combinations are obtained when compared with respect to all possible forecast combinations constructed from a set of time-series model forecasts. The results from the empirical analysis indicate that the worst case scenario with SDE weights performs the worst when compared with other forecast selection and combination methods at all sections of the forecast error distribution for all horizons. On the other hand, optimal forecast combination with SDE weights perform better or equally well when compared with standard forecast selection and combination models in the literature. We have constructed the optimal (worst) forecast combination for different forecast horizons at different percentiles of the forecast error distribution for weekly Japanese yen/U.S. dollar and U.S. dollar/Great Britain pound foreign exchange rate forecasts. We have found that the RW model is the main contributor to the worst forecast combinations across all sections of the forecast error distribution in both applications. Conversely, the NNETTS and SETAR models, on average, contribute more in the optimal forecast combinations with SDE weights for both the weekly Japanese yen/U.S. dollar and the U.S. dollar/Great Britain pound exchange rate forecasts. However, the NNETTS and SETAR models contribute the most at different parts of the forecast error distribution. Overall, there is also agreement in both applications that the ARMA, MS, RW and LSTAR models contribute more to the optimal forecast combination at some parts of the forecast error distribution, albeit only for some forecast horizons.

In summary, for the majority of the cases considered, forecast combinations with SDE weights perform better than median forecasts, forecasts from the model that is chosen with AIC, BIC, predictive least squares, and forecast combination with equal, Bates and Granger, AIC, and BIC weights. However, there are also few cases where some other forecast selection and combination model may perform equally well at some parts of the forecast error distribution. In only one occasion, the forecast combination with Bates and Granger weights performs better than that with SDE weights. Our findings suggest a partial support for the forecast combination puzzle

because there are many other forecast combinations that are less efficient than the average forecast combination, and the equally weighted model yields the optimal combination in some cases. We should mention that we only applied SDE analysis to two specific data sets with a given number (seven) of time-series models and, as such, our results on the optimality of the forecast combination does not generalize beyond the scope of the applications at hand. However, the SDE methodology can offer a useful way of assessing the optimality of forecast combinations by using all of the information available in the entire forecast error distribution and not merely the first two moments, as typically observed in the literature.

References

[1] Agliardi, E., Agliardi, R., Pinar, M. Stengos, T., and N. Topaloglou (2012) A new country risk index for emerging markets: A stochastic dominance approach, Journal of Empirical Finance, 19(5), 741-761.

[2] Al-Qassam, M.S., and J.A. Lane (1989). Forecasting exponential autoregressive models of order 1. Journal of Time Series Analysis, 10(2), 95-113.

[3] Abadie, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. Journal of the American Statistical Association, 97(457), 284-292.

[4] Aiolfi, M., Capistrán, C., and A. Timmermann (2011). Forecast combination. In: Clements, M.P., Hendry, D.F. (eds.), The Oxford handbook of economic forecasting. Oxford University Press, Oxford.

[5] Anderson, G. (1996). Nonparametric tests of stochastic dominance in income distributions. Econometrica, 64(5), 1183-1193.

[6] Barrett, G.F, and S.G. Donald (2003). Consistent tests for stochastic dominance. Econometrica, 71(1), 71-104.

[7] Bates, J.M., and C.W.J. Granger (1969). The combination of forecasts. Operations Research Quarterly, 20(4), 451-468.

[8] Beach, C., and R. Davidson (1983). Distribution-free statistical inference with Lorenz curves and income shares. Review of Economic Studies, 50(4), 723-735.

[9] Burnham, K.P., and D.R. Anderson (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd ed., Springer, New York.

[10] Carlstein, E. (1986). The use of subseries methods for estimating the variance of a general statistic from a stationary time series. Annals of Statistics, 14(3), 1171–1179.

[11] Clemen, R.T. (1989). Combining forecasts: A review and annotated bibliography. International Journal of Forecasting, 5(4), 559-581.

[12] Clements, M.P., and D.F. Hendry (1998). Forecasting Economic Time Series. Cambridge University Press, Cambridge, U.K.

[13] Clements, M.P., and D.F. Hendry (1999). Forecasting Non-stationary Economic Time Series. MIT Press, Cambridge, MA.

[14] Clements, M.P., and D.F. Hendry (2006). Forecasting with breaks. In: Elliott, G., Granger, C.W.J., Timmermann, A. (eds), Handbook of Economic Forecasting. North-Holland, Amsterdam, pp. 605-658.

[15] Davidson, R., and J.-Y. Duclos (2000). Statistical inference for stochastic dominance and for the measurement of poverty and inequality. Econometrica, 68(6), 1435-1464.

[16] Diebold, F.X., and P. Pauly (1987). Structural change and the combination of forecasts. Journal of Forecasting, 6(1), 21-40.

[17] De Gooijer, J.G., and P.T. De Bruin (1998) On forecasting SETAR processes. Statistics and Probability Letters, 37(1), 7-14.

[18] Drechsel, K., L. Maurin (2010) Flow of Conjunctural Information and Forecast of Euro Area Economic Activity, Journal of Forecasting, 30, 336354.

[19] Elliott, G., and A. Timmermann (2004). Optimal forecast combinations under general loss functions and forecast error distributions. Journal of Econometrics, 122(1), 47-79.

[20] Elliott, G., and A. Timmermann (2005). Optimal forecast combination under regime switching. International Economic Review, 46(4), 1081-1102.

[21] Elliott, G., and A. Timmermann (2008). Economic forecasting. Journal of Economic Literature, 46(1), 3-56.

[22] Fernández C., Ley E., and M.F. Steel (2001). Model uncertainty in cross-country growth regressions. Journal of Applied Econometrics 16(5), 563-576.

[23] Franses, P., and D. Van Dijk (2000). Nonlinear time series models in empirical finance. Cambridge University Press, Cambridge.

[24] Geweke, J., and G. Amisano (2011). Optimal prediction pools. Journal of Econometrics, 164(1), 130-141.

[25] Geweke, J., and G. Amisano (2012). Prediction with misspecified models. American Economic Review, 102(3), 482-486.

[26] Granger, C.W.J. (1999). Outline of forecast theory using generalized cost functions. Spanish Economic Review, 1(2), 161-173.

[27] Hamilton, J. (1994). Time series analysis. Princeton University Press, Princeton.

[28] Hendry, D.F., and M.P. Clements (2004). Pooling of forecasts. Econometrics Journal, 7(1), 1-31.

[29] Huang, H., and T.-H. Lee (2010). To combine forecasts or to combine information? Econometric Reviews, 29(5), 534-570.

[30] Hyndmana, R.J., and A.B. Koehlerb (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679-688.

[31] Kock, A.B., and T. Terasvirta (2011). Forecasting with nonlinear time series models. In: Clements, M.P., Hendry, D.F. (eds.), Oxford handbook of economic forecasting. Oxford University Press, Oxford, pp. 61-87.

[32] Kolassa, S. (2011). Combining exponential smoothing forecasts using Akaike weights. International Journal of Forecasting, 27(2), 238-251.

[33] Lin, J.L., and C.W.J. Granger (1994). Forecasting from non-linear models in practice. Journal of Forecasting, 13(1), 1-9.

[34] Linton, O., Maasoumi, E., and Y.-J. Whang (2005). Consistent testing for stochastic dominance under general sampling schemes. Review of Economic Studies, 72(3), 735-765.

[35] Marcellino, M., Stock, J.H., and M.W. Watson (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. Journal of Econometrics, 135(1-2), 499-526.

[36] Pinar, M., Stengos, T., and N. Topaloglou (2013) Measuring human development: a stochastic dominance approach. Journal of Economic Growth, 18(1), 69-108.

[37] Raftery, A.E. (1995). Bayesian model selection in social research. Sociological Methodology, 25, 111-163.

[38] Rio, E. (2000). Theorie Asymptotique des Processus Aleatoires Faiblement Dependants. Springer-Verlag, Berlin.

[39] Rissanen, J. (1986). Order estimation by accumulated prediction errors. Journal of Applied Probability, 23A, 55-61.

[40] Sala-i-Martin, X., Doppelhofer, G., and R.I. Miller (2004). Determinants of long-term growth: a Bayesian averaging of classical estimates (BACE) approach. American Economic Review, 94(4), 813-835.

[41] Scaillet, O., and N. Topaloglou (2010). Testing for stochastic dominance efficiency. Journal of Business and Economic Statistics, 28(1), 169-180.

[42] Smith, J., and K.F. Wallis (2009). A simple explanation of the forecast combination puzzle. Oxford Bulletin of Economics and Statistics, 71(3), 331-355.

[43] Stock, J.H., and M.W. Watson (1999a). Forecasting inflation. Journal of Monetary Economics, 44(2), 293-335.

[44] Stock, J.H., and M.W. Watson (1999b). A comparison of linear and nonlinear models for forecasting macroeconomic time series. In: Engle, R.F., White, H. (eds.), Cointegration, Causality and Forecasting. Oxford University Press, Oxford, pp. 1-44.

[45] Stock, J.H., and M.W. Watson (2004) Combination forecasts of output growth in a seven-country data set. Journal of Forecasting, 23(6), 405-430.

[46] Swanson, N.R., and T. Zeng (2001) Choosing Among Competing Econometric Forecasts: Regression-based Forecast Combination Using Model Selection. Journal of Forecasting, 20, 425-440.

[47] Terasvirta, T. (2006). Forecasting economic variables with nonlinear models. In: Elliott, G., Granger, C.W.J., Timmermann, A. (eds), Handbook of Economic Forecasting, Volume 1. Elsevier, Oxford, pp. 413-457.

[48] Timmermann A. (2006). Forecast combinations. In: Elliott, G., Granger, C.W.J., Timmermann, A. (eds), Handbook of Economic Forecasting. North-Holland, Amsterdam, pp. 135-196.

[49] Van Dijk, D., Franses, P.H., Clements, M.P., and J. Smith (2003). On SETAR non-linearity and forecasting. Journal of Forecasting 22(5), 359-375.

[50] White, H. (2006). Approximate nonlinear forecasting methods. In: Elliott, G., Granger, C.W.J., Timmermann, A. (eds), Handbook of Economic Forecasting, Volume 1. Elsevier, Oxford, pp. 459-512.

Tables

Table 1: Optimal and worst forecast combinations (Japanese yen/U.S. dollar exchange rates)

Forecast horizon	**Percentile**	**Forecast error level (average forecast comb.)**	**WEIGHTS**													
			AR		**ARMA**		**LSTAR**		**MS**		**NNETTS**		**RW**		**SETAR**	
			Optimal	**Worst**	**Optimal**	**Worst**	**Optimal**	**Worst**	**Optimal**	**Worst**	**Optimal**	**Worst**	**Optimal**	**Worst**	**Optimal**	**Worst**
h=1	50th	0.0104	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0451	0.0437	0.0000	0.9563	0.9549	0.0000
	75th	0.0172	0.0000	0.0000	0.0000	0.0035	0.0000	0.0000	0.0000	0.0000	0.8408	0.0000	0.0000	0.9965	0.1592	0.0000
	95th	0.0318	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1153	0.0000	0.0336	1.0000	0.8511	0.0000
h=27	50th	0.0113	0.0000	0.0000	0.0000	0.0550	0.0000	0.0000	0.0000	0.0000	0.0297	0.0000	0.1912	0.9450	0.7791	0.0000
	75th	0.0178	0.0000	0.0000	0.0000	0.0000	0.0286	0.0000	0.0000	0.0000	0.9714	0.0000	0.0000	0.9592	0.0000	0.0408
	95th	0.0318	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	1.0000	NA	0.0000
h=53	50th	0.0115	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4732	0.0000	0.2607	1.0000	0.2661	0.0000
	75th	0.0173	0.0000	0.0000	0.1246	0.0163	0.6322	0.0000	0.0000	0.0000	0.2165	0.0000	0.0267	0.9837	0.0000	0.0000
	95th	0.0309	0.0000	0.0000	0.0000	0.0118	0.0000	0.0000	0.0000	0.0000	0.8015	0.0000	0.1985	0.9882	0.0000	0.0000
h=105	50th	0.0106	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.8254	0.0532	0.1746	0.9468	0.0000	0.0000
	75th	0.0162	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7690	0.0326	0.2310	0.9674	0.0000	0.0000
	95th	0.0278	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	1.0000	NA	0.0000

Table 2: Number of forecast errors less than a given forecast error level (Japanese yen/U.S. dollar exchange rates)

Forecast horizon	Percentile	Forecast error level (at EW)	EW	Median	AIC	BIC	PLS	AIC weights	BIC weights	Bates-Granger weights	SDE Worst	SDE Best
h=1	50th	0.0104	104	102	100	100	100	100	100	102	81	**107**
	75th	0.0172	156	158	157	157	157	157	157	154	128	**159**
	95th	0.0318	197	198	198	198	198	198	198	198	179	**200**
h=27	50th	0.0113	91	85	85	85	88	85	85	87	69	**94**
	75th	0.0178	136	138	138	138	133	138	138	135	104	**139**
	95th	0.0318	**173**	172	172	172	172	172	172	**173**	154	**173**
h=53	50th	0.0115	78	74	74	74	75	72	73	74	57	**79**
	75th	0.0173	117	115	114	114	112	113	114	114	90	**118**
	95th	0.0309	148	147	147	147	147	147	147	147	130	**149**
h=105	50th	0.0106	52	50	50	50	50	50	50	52	44	**53**
	75th	0.0173	78	76	76	76	76	76	76	75	59	**79**
	95th	0.0278	**99**	**99**	**99**	**99**	**99**	**99**	**99**	**99**	86	**99**

Table 3: Average weights of the worst and optimal forecast combinations for the entire distribution (Japanese yen/U.S. dollar exchange rates)

Forecast horizon	Forecast combination	AR	ARMA	LSTAR	MS	NNETTS	RW	SETAR
h=1	Optimal	0.0137	0.0394	0.0021	0.1195	0.5707	0.0078	0.2468
	Worst	0.0026	0.0220	0.0000	0.0049	0.0899	0.7704	0.1102
h=27	Optimal	0.0058	0.1201	0.1558	0.0051	0.4234	0.0154	0.2744
	Worst	0.0008	0.0768	0.0315	0.0001	0.1092	0.7545	0.0271
h=53	Optimal	0.0408	0.0907	0.1128	0.0021	0.3904	0.0153	0.3479
	Worst	0.0000	0.0403	0.0071	0.0000	0.0000	0.9526	0.0000
h=105	Optimal	0.0102	0.0490	0.0032	0.0310	0.6241	0.0215	0.2610
	Worst	0.0000	0.0000	0.0000	0.0000	0.0418	0.9582	0.0000

Table 4: Optimal and worst forecast combinations (U.S. dollar/Great Britain pound exchange rates)

Forecast horizon	**Percentile**	**Forecast error level (average forecast comb.)**	**WEIGHTS**													
			AR		**ARMA**		**LSTAR**		**MS**		**NNETTS**		**RW**		**SETAR**	
			Optimal	**Worst**	**Optimal**	**Worst**	**Optimal**	**Worst**	**Optimal**	**Worst**	**Optimal**	**Worst**	**Optimal**	**Worst**	**Optimal**	**Worst**
h=1	50th	0.0097	0.0000	0.0000	0.0007	0.0000	0.0000	0.0000	0.0000	0.0000	0.0587	0.0063	0.0000	0.9937	0.9406	0.0000
	75th	0.0176	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0387	0.0084	0.0819	0.9916	0.8794	0.0000
	95th	0.0386	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.6931	0.0279	0.3069	0.9721	0.0000	0.0000
h=27	50th	0.0097	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0333	0.0000	0.9667	1.0000	0.0000
	75th	0.0191	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9312	0.0257	0.0493	0.9743	0.0195	0.0000
	95th	0.0403	0.0000	0.0000	0.2276	0.0000	0.0000	0.0000	0.0000	0.0000	0.5592	0.0000	0.2132	1.0000	0.0000	0.0000
h=53	50th	0.0104	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0842	0.0156	0.0140	0.9844	0.9018	0.0000
	75th	0.0201	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9873	0.0222	0.0000	0.9778	0.0127	0.0000
	95th	0.0405	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	0.0000	NA	1.0000	NA	0.0000
h=105	50th	0.0112	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2741	0.0000	0.6371	0.0468	0.0888	0.9532	0.0000	0.0000
	75th	0.0194	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2741	0.0000	0.6371	0.0206	0.0888	0.9794	0.0000	0.0000
	95th	0.0395	0.0000	0.0000	0.0454	0.0000	0.0000	0.0000	0.0000	0.0000	0.6064	0.0000	0.3482	1.0000	0.0000	0.0000

Table 5: Number of forecast errors less than a given forecast error level (U.S. dollar/Great Britain pound exchange rates)

Forecast horizon	Percentile	Forecast error level (at EW)	EW	Median	AIC	BIC	PLS	AIC weights	BIC weights	Bates - Granger weights	SDE Worst	SDE Best
h=1	50th	0.0097	104	108	108	108	111	108	108	104	68	**112**
	75th	0.0176	156	155	156	156	154	155	155	155	119	**159**
	95th	0.0386	197	197	197	197	197	197	197	196	188	**199**
h=27	50th	0.0097	91	88	88	88	**92**	88	88	90	58	**92**
	75th	0.0191	136	138	138	138	139	138	138	138	103	**141**
	95th	0.0403	173	172	172	172	172	172	172	172	158	**175**
h=53	50th	0.0104	78	75	75	75	79	75	75	**81**	48	79
	75th	0.0201	117	120	**121**	**121**	117	**121**	**121**	119	83	**121**
	95th	0.0405	**148**	146	146	146	146	146	146	146	130	**148**
h=105	50th	0.0112	52	54	54	54	52	52	54	54	42	**55**
	75th	0.0194	78	78	78	78	77	77	78	77	68	**79**
	95th	0.0395	99	99	99	99	99	99	99	99	89	**100**

Table 6: Average weights of the worst and optimal forecast combinations for the entire distribution (U.S. dollar/Great Britain pound exchange rates)

Forecast horizon	Forecast combination	AR	ARMA	LSTAR	MS	NNETTS	RW	SETAR
h=1	Optimal	0.0162	0.1330	0.0000	0.0586	0.4625	0.0140	0.3157
	Worst	0.0000	0.0000	0.0000	0.0000	0.0016	0.9984	0.0000
h=27	Optimal	0.0124	0.1008	0.0063	0.0409	0.5086	0.0116	0.3194
	Worst	0.0009	0.0427	0.0000	0.0051	0.0996	0.7434	0.1083
h=53	Optimal	0.0784	0.1060	0.0000	0.0358	0.3981	0.0113	0.3704
	Worst	0.0000	0.0000	0.0000	0.0000	0.0230	0.9770	0.0000
h=105	Optimal	0.0119	0.1897	0.0909	0.0204	0.5452	0.0221	0.1198
	Worst	0.0000	0.0000	0.0000	0.0000	0.1273	0.8727	0.0000

Appendix : Simulating p-values

Block Bootstrap Methods

In this appendix, we describe practical ways to compute p-values for testing stochastic dominance efficiency at any order by using block bootstrap methods and discuss the theoretical justification for these methods. Block bootstrap methods apply the nonparametric i.i.d. bootstrap to a time series context (see Barrett and Donald (2003) and Abadie (2002) for the use of the nonparametric i.i.d. bootstrap in stochastic dominance tests). They are based on "blocking" arguments which data are divided into blocks; these blocks, rather than the individual data, are resampled to mimic the time-dependent structure of the original data. We focus on block bootstrapping because we deal with moderate sample sizes in the empirical applications and wish to exploit the full sample information.

Let b, l denote integers such that $N_f = bl$. The non-overlapping rule (Carlstein, 1986) merely asks that the data to be divided into b disjoint blocks, the k-th being $\boldsymbol{B}_k = ((y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t})'_{(k-1)l+1}, ..., (y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t})'_{kl})'$, with $k \in \{1, ..., b\}$. The block bootstrap method requires that we choose blocks $\boldsymbol{B}_1^*, ..., \boldsymbol{B}_b^*$ by resampling randomly, with replacement, from the set of non-overlapping blocks. If $\boldsymbol{B}_i^* = ((y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t})^{*\prime}_{i1}, ..., (y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t})^{*\prime}_{il})'$, a block bootstrap sample $\{(y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t})^*_{t+h};\ \ N_f = 1, ..., N_f\}$ is made of $\{(y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t})^*_{11}, ..., (y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t})^*_{1l}, (y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t})^*_{21}, ..., (y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t})^*_{2l}, ..., (y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t})^*_{b1}, ..., (y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t})^*_{bl}\}$, and we let $\hat{F}^*$ denote its empirical distribution.

Let us define $p_j^* := P[S_j^* > \hat{S}_j]$, where S_j^* is the test statistic corresponding to each bootstrap sample. Then, the block bootstrap method is justified by the following statement (and the proof is given by ST).

Proposition 1 *Assuming that $\alpha < 1/2$, a test for SDE_j based on the rule*

$$\text{“ reject } H_0^j \quad \text{if} \quad p_j^* < \alpha \text{ ”},$$

satisfies the following

$$\lim P[\text{reject } H_0^j] \leq \alpha \qquad \text{if} H_0^j \text{ is true,}$$
$$\lim P[\text{reject } H_0^j] = 1 \qquad \text{if} H_0^j \text{ is false.}$$

In practice, we need to use Monte Carlo methods to approximate the probability. The p-value is simply approximated by $\tilde{p}_j = \dfrac{1}{R}\sum_{r=1}^{R} \mathbb{I}\left\{\widetilde{S}_{j,r} > \widehat{S}_j\right\}$, where the average is made based on R replications. The replication number can be chosen to make the approximations as accurate as we desire given time and computer constraints.

Mathematical formulation of the test statistics

The test statistic $\hat{S}_1$ for first-order stochastic dominance efficiency is derived using mixed integer programming formulations. The following is the full formulation of the model:

$$\max_{\mathbf{z},\boldsymbol{\lambda}} \hat{S}_1 = \sqrt{N_f}\frac{1}{N_f}\sum_{N_f=1}^{N_f}(L_{N_f} - W_{N_f}) \tag{17}$$

$$\text{s.t.} M(L_{N_f} - 1) \leq z - (y_{t+h} - \boldsymbol{\tau}^{'}\widehat{\boldsymbol{y}}_{t+h,t}) \leq ML_{N_f}, \qquad \forall\, t \tag{18}$$

$$M(W_{N_f} - 1) \leq z - (y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}) \leq MW_{N_f}, \qquad \forall\, t \tag{19}$$

$$\boldsymbol{e}'\boldsymbol{\lambda} = 1, \tag{20}$$

$$\boldsymbol{\lambda} \geq 0, \tag{21}$$

$$W_{N_f} \in \{0,1\},\ L_{N_f} \in \{0,1\}, \qquad \forall\, t \tag{22}$$

with M being a large constant.

The model is a mixed integer program maximizing the distance between the two binary variables, $\frac{1}{N_f}\sum_{N_f=1}^{N_f} L_{N_f}$ and $\frac{1}{N_f}\sum_{N_f=1}^{N_f} W_{N_f}$, which represent $G(z,\boldsymbol{\tau};\hat{F})$ and $G(z,\boldsymbol{\lambda};\hat{F})$, respectively (the empirical cdf of the forecast errors with the forecast combination are $\boldsymbol{\tau}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$ and $\boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$, respectively, at forecast error level z). According to inequality (14), L_{N_f} equals 1 for each scenario of realization factors N_f for which $z \geq (y_{t+h} - \boldsymbol{\tau}^{'}\widehat{\boldsymbol{y}}_{t+h,t})$ and equals 0 otherwise. Analogously, inequality (15) ensures that W_{N_f} equals 1 for each scenario for which $z \geq (y_{t+h} - \boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t})$. Equation (16) defines the sum of all forecast combination weights to be unity, while inequality (17) disallows for negative weights.

This formulation allows us to test the dominance of the equally weighted forecast combination, $\boldsymbol{\tau}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$, over any potential linear forecast combination, $\boldsymbol{\lambda}^{'}\widehat{\boldsymbol{y}}_{t+h,t}$, of the forecasts based on time-series models. When some of the variables are binary, corresponding to mixed integer programming, the problem becomes non-polynomial (NP)-complete (i.e., formally intractable). The problem can be reformulated to reduce the solving time and to obtain a tractable formulation (see section 4.1 of ST, for the derivation of this formulation and details on its practical implementation).