Jensen's Inequality and the Success of Linear Prediction Pools*

Fabian Krüger†
University of Konstanz

May 22, 2012

Preliminary Draft – Comments Welcome

Abstract

Combinations of density forecasts (so-called "prediction pools") are a natural way to account for model uncertainty. Using the log score as an evaluation criterion, a number of recent papers have found that (linear) prediction pools produce good forecasts of key economic variables. The present paper provides evidence that the success of linear pools is not restricted to the log score, but carries over to the quadratic and continuous ranked probability scores. These scoring rules possess several attractive features and should at least be considered as viable alternatives to the log score. I show that under all three scoring rules, Jensen's inequality sets a lower bound on the success of linear pools, relative to the models which constitute the pool. Hence, pooling serves as a hedge against picking a wrong individual model. Monte Carlo evidence suggests that this hedge can be available at low cost, in the sense that (misspecified) pools often perform similar to the true model. An application to US macro data demonstrates the empirical implications of these findings.

Keywords: Density Forecasting, Forecast Combination, Scoring Rules
JEL Classification: C52

*I thank Tilmann Gneiting, Ingmar Nolte, Winfried Pohlmeier and Peter Schanbacher for helpful comments. The usual disclaimer applies.

†Department of Economics, Universitätsstraße 1, D-78462 Konstanz, Germany. Phone: +49-7531-88-3753, email: Fabian.Krueger@uni-konstanz.de. Financial support from the Fritz Thyssen foundation is gratefully acknowledged.

1 Introduction

Most economic decisions require probabilistic forecasts which provide information on a range of possible future scenarios. This stands in contrast to traditional point forecasts which indicate a single "expected" outcome. The need for probabilistic forecasts raises two questions: First, what is a good probabilistic forecast? Second, how can it be constructed? Building upon earlier work in meteorology and other fields, these two questions have recently received much attention from econometricians.

Similar to the literature on combinations of point forecasts (Timmermann, 2006), combining probabilistic forecasts is a natural idea. Rather than choosing a particular model, combinations average over several available models. Following a proposal by Wallis (2005), combinations of probabilistic forecasts typically take the form of linear prediction pools. Using the log score (the out-of-sample log likelihood) as an evaluation criterion, a number of studies have demonstrated that linear pools provide good density forecasts of key economic variables, where "good" is meant relative to the individual models which enter the pool. See e.g. Hall and Mitchell (2007), Jore, Mitchell, and Vahey (2010), Kascha and Ravazzolo (2010), Bache, Jore, Mitchell, and Vahey (2011) and Geweke and Amisano (2011).

It should be stressed, however, that the log score is by far not the only available scoring rule which specifies the forecast user's utility from a density forecast $f(\cdot)$ and an outcome y that materializes. Gneiting and Raftery (2007) provide a comprehensive review of alternative scoring rules. All of the common rules are proper, i.e. they are maximized in expectation by the (unknown) true model. However, they may well give different rankings of alternative misspecified models. Hence, it seems important to analyze whether the positive results on linear prediction pools hold true under scoring rules other than the log score.

Focusing on discrete outcome variables, a number of studies (e.g. Boero, Smith, and

Wallis, 2011 and Clements and Harvey, 2011) have considered the performance of linear prediction pools under alternative scoring rules. By contrast, evidence for continuous outcome variables is almost entirely missing from the economic literature, with Gneiting and Thorarinsdottir (2010) being a notable exception. This lack of results is a shortcoming of the extant literature, since i) most economic variables are treated as continuous, and ii) the papers focussing on discrete outcomes do not cover the forecasting models typically used in macroeconomics.

This paper analyzes the performance of linear prediction pools for continuous variables, under two scoring rules: The quadratic score (Brier, 1950) and the continuous ranked probability score (Matheson and Winkler, 1976). For a number of conceptual and practical reasons detailed below, these scoring rules should at least be considered as viable alternatives to the log score. They have been used extensively in meteorology, but not in economics. Throughout, I include the log score as a reference case. The paper is divided into three parts.

Firstly, I show that under all three scoring rules, Jensen's inequality sets a lower bound on the success of linear pools, relative to the models which constitute the pool. The inequality holds for all realizing values of the predictand, and thus also in expectation over any (unknown) true model. This makes pooling very attractive from an ex ante perspective.

Secondly, I use a Monte Carlo study to analyze the performance of linear pools in a dynamic setting. The individual models to be combined are several autoregressive models calibrated to match monthly US macro variables. I show that even if one of the individual models coincides with the true data-generating process, a (misspecified) linear pool often comes close to the performance of the true model. Tests for equal predictive ability (Giacomini and White, 2006) tend to have little power in discriminating between the linear pool and the true model, even for large evaluation samples.

Thirdly, I provide empirical results for US macro variables and different vector autoregressive specifications. I find that under all three scoring rules and for most variables and forecast horizons, equally weighted linear pools perform very well relative to the individual models. In half of the forecast comparisons, equally weighted pools even outperform *all* individual models on average over the evaluation sample. The intuition for this result is that, by Jensen's inequality, the equally weighted pool performs relatively well *for each point in the evaluation sample*. In contrast, even good individual models fail from time to time.

In summary, my analysis shows that the success of linear pools is not restricted to the log score, but carries over to two other major scoring rules. It appears that little can be lost but much can be gained by combining a set of density forecasts, rather than picking a single forecast.

The rest of this paper is organized as follows. Section 2 briefly reviews the scoring rules considered in this paper. Sections 3, 4 and 5 present analytical, simulation and empirical results, respectively. Section 6 concludes. All proofs are relegated to the appendix.

2 Scoring Rules

This section presents the three scoring rules considered in the following; detailed surveys on the topic are provided by Winkler (1996) and Gneiting and Raftery (2007).

Suppose a forecaster issues a probability density function (p.d.f.) $f(\cdot)$ for a real-valued random variables Y, and an outcome $y \in \mathbb{R}$ materializes. It is not immediately clear how to judge whether the forecast $f(\cdot)$ was "good" or "bad", since the true density of Y (say, $f_0(\cdot)$) is unobservable even ex post. All that is observed is a single draw $Y = y$ from this density.

Scoring rules assign a real value based on y and $f(\cdot)$; I take them to be positively oriented (the larger the better). In the following, I consider three scoring rules: The logarithmic ("log"), quadratic, and continuous ranked probability scores:

$$LS(y, f(\cdot)) = \ln f(y), \tag{1}$$

$$QS(y, f(\cdot)) = 2f(y) - \int f^2(z)dz, \tag{2}$$

$$CRPS(y, f(\cdot)) = -\int (F(z) - \mathbb{I}_{\{z \geq y\}})^2 dz, \tag{3}$$

where $F(\cdot)$ is the cumulative density function (c.d.f.) implied by $f(\cdot)$ and $\mathbb{I}_{\{A\}}$ is the indicator function of the event A.

All three scoring rules are strictly proper, that is, a forecaster who knows the true p.d.f. $f_0(\cdot)$ can maximize her expected score by actually stating $f_0(\cdot)$, instead of another density $f(\cdot)$:

$$\int S(z, f_0(\cdot)) f_0(z)dz > \int S(z, f(\cdot)) f_0(z)dz, \tag{4}$$

for any density forecast $f(\cdot) \neq f_0(\cdot)$, where $S(\cdot, \cdot)$ is one of the three scoring rules above. Note that for each rule, Equation (4) is based on measure-theoretic regularity conditions which are detailed in Gneiting and Raftery (2007).

Propriety is typically considered a minimal criterion for a "reasonable" scoring rule. The fact that the three rules are proper implies that in expectation, they all favor the (unknown) true model over any competitor. However, they may well give different rankings of alternative misspecified models. In practice, all models are misspecified; hence nothing says that empirical forecast comparisons should be robust to the choice of scoring rule. This motivates my analysis of linear prediction pools under scoring rules other than the log score.

As stated earlier, virtually all econometric studies on probabilistic forecasting of continuous variables use the log score criterion proposed by Good (1952). The popularity of the

log score may be due to its close relation to (log) likelihood and the Kullback and Leibler (1951) divergence. From Equations (1) to (3), an important difference between the log score and the other two scoring rules becomes apparent: The log score is "local". This means that the assigned score depends only on the predictive density at the value $Y = y$ that actually materializes, and not on the density at other values. Locality often makes the log score simpler to handle than the other two scoring rules, which involve potentially complicated integrals. An important practical drawback of the log score is its lack of robustness which results from the fact that $\ln(f(y)) \to -\infty$ as $f(y) \to 0$. Hence in the tails of $f(\cdot)$, small differences in the outcome y – which may be due to rounding, data revisions, etc. – can lead to drastic differences in the assigned log score (Selten, 1998).

The quadratic score (QS) in Equation (2) is a continuous version of the Brier (1950) score for probability forecasts of discrete events, which has been used extensively across many fields. Like its discrete counterpart, QS in (2) is neutral, i.e. the expected score of a forecast $f(\cdot)$ if $f_0(\cdot)$ is the true density is the same as the expected score of forecast $f_0(\cdot)$ if $f(\cdot)$ is true (Gneiting and Raftery, 2007, p. 365). Selten (1998, p. 54) argues in favor of this property, and points out some other attractive features of QS in a discrete setting. In particular, QS does not share the log score's sensitivity to tail events.

The cumulative ranked probability score ($CRPS$) in Equation (3) has been proposed by Matheson and Winkler (1976); unlike the two other rules, it is in terms of a predictive c.d.f., rather than a p.d.f. It is very popular in meteorology (Hersbach, 2000) and has recently been employed in econometrics (Gneiting and Thorarinsdottir, 2010; Gneiting and Ranjan, 2011). Unlike the log and quadratic scores, the $CRPS$ rewards predictive densities which have probability mass *near* but not *at* the realizing value y (i.e., the $CRPS$ is "sensitive to distance"). This can be an important benefit in the presence of noisy data. Furthermore, Hersbach (2000) remarks that the $CRPS$ reduces to the absolute error if the predictive "density" is a point forecast. Hence the $CRPS$ can in principle be used to compare point and density forecasts. A practical drawback of the

$CRPS$ is that for non-Gaussian models, the integral in (3) is typically not available in closed form.

3 Jensen's Inequality and the Success of Linear Prediction Pools

In this section, I consider a one-shot scenario in which a single outcome of a random variable Y is to be predicted. I consider linear prediction pools of the form

$$f_c(Y) = \sum_{i=1}^{n} \omega_i f_i(Y), \tag{5}$$

where $f_i(Y),\ i = 1, \ldots, n$ are n predictive densities for a random variable Y, and the weights satisfy $\omega_i \geq 0\ \forall\ i$ and $\sum_{i=1}^{n} \omega_i = 1$. The mean μ_c and variance σ_c^2 of the linear pool are given by

$$\begin{aligned} \mu_c &= \sum_{i=1}^{n} \omega_i \mu_i, \\ \sigma_c^2 &= \sum_{i=1}^{n} \omega_i \sigma_i^2 + \sum_{i=1}^{n} \omega_i (\mu_i - \mu_c)^2, \end{aligned}$$

where μ_i and σ_i^2 are the mean and variance of model i.

Although linear pools have long been known as a method for aggregating density functions (Stone, 1961), it appears that the econometric use of linear pools has first been proposed by Wallis (2005). Geweke and Amisano (2011) provide a detailed analysis of linear prediction pools under the log score.

Equation (6) implies that the variance of the pool is generally larger than the minimal variance across the n individual models: $\sigma_c^2 \geq \min_{i \in \{1,\ldots,n\}} \sigma_i^2$. Furthermore, if the means

of the individual models differ substantially ($|\mu_i - \mu_j|$ "large" for some i, j), the variance of the pool can even exceed the variances of *all* individual models. Taken together, this implies that linear pooling will often produce very dispersed predictive densities.

McNees (1992) points out that in terms of the squared prediction error, the equally weighted mean of n point forecasts necessarily performs better than the average of the n individual models. This result is a simple arithmetical consequence of Jensen's inequality. Manski (2010) emphasizes the power of this statement, and argues that it has been under-appreciated by the extensive literature which analyzes the success of forecast combinations. Perhaps surprisingly, results which are similar to those of McNees (1992) apply to combinations of density forecasts, under all three scoring rules considered here.

Proposition 3.1. *Consider a linear pool as defined in Equation (5) and the three scoring rules defined in (1) to (3). Then, if an outcome $Y = y$ materializes,*

$$LS(y, \sum_{i=1}^{n} \omega_i f_i(\cdot)) \geq \sum_{i=1}^{n} \omega_i LS(y, f_i(\cdot)), \tag{6}$$

$$QS(y, \sum_{i=1}^{n} \omega_i f_i(\cdot)) \geq \sum_{i=1}^{n} \omega_i QS(y, f_i(\cdot)), \tag{7}$$

$$CRPS(y, \sum_{i=1}^{n} \omega_i f_i(\cdot)) \geq \sum_{i=1}^{n} \omega_i CRPS(y, f_i(\cdot)). \tag{8}$$

Equation (6) is a slightly more general version of the result in Kascha and Ravazzolo (2010, p. 237). To the best of my knowledge, Equations (7) and (8) are novel to the literature.

Proposition 3.1 gives a lower bound on the score of the prediction pool, for any realizing observation y and under any of the three scoring rules. This lower bound depends on the scores attained by the n individual models and the weights ω_i. For the special case $n = 2$, the inequalities in the proposition boil down to a definition of concavity when the

argument is function-valued; c.f. Equation (3) of Gneiting and Raftery (2007). Hence for given y, the three scoring rules $S(y, f(\cdot))$ are concave functions of $f(\cdot)$.

The next result shows that the nature of the bounds from Proposition 3.1 is different for LS than for QS and $CRPS$.

Corollary 3.1. *Define* $\Delta_{LS} = LS(y, \sum_{i=1}^n \omega_i f_i(\cdot)) - \sum_{i=1}^n \omega_i LS(y, f_i(\cdot))$, *and similarly for* QS *and* $CRPS$. *Then,*

$$\Delta_{LS} = \ln\left\{\frac{\sum_{i=1}^n \omega_i f_i(y)}{\prod_{i=1}^n f_i(y)^{\omega_i}}\right\} \geq 0, \tag{9}$$

$$\Delta_{QS} = \int \left\{\sum_{i=1}^n \omega_i f_i^2(z) - (\sum_{i=1}^n \omega_i f_i(z))^2\right\} dz \geq 0, \tag{10}$$

$$\Delta_{CRPS} = \int \left\{\sum_{i=1}^n \omega_i F_i^2(z) - (\sum_{i=1}^n \omega_i F_i(z))^2\right\} dz \geq 0. \tag{11}$$

The term Δ_{LS} can be interpreted as a normalized performance measure for linear pools, which accounts for the tautological lower bound from Equation (6). It can be seen from (9) that Δ_{LS} depends on the realizing outcome y. It is easy to construct examples in which the term takes arbitrarily large values.

Example 3.1. *Consider* $n = 2$ *Gaussian densities with means* μ_i *and variances* σ_i^2 *(*$i = 1, 2$*), as well as their linear pool with weights* $(\omega, 1-\omega)$, *where* $0 < \omega < 1$. *Then if* $\sigma_1^2 \neq \sigma_2^2$, $\Delta_{LS} \to \infty$ *as* $y \to \infty$ *or* $y \to -\infty$.

The behavior of Δ_{LS} stands in contrast to Δ_{QS} and Δ_{CRPS} in Equations (10) and (11). Both terms do not depend on y, but are deterministic functions of the component models $\{f_i(\cdot)\}_{i=1}^n$ and the weights $\{\omega_i\}_{i=1}^n$. Hence Corollary 3.1 implies that the benefits of pooling are stochastic and unbounded under LS, whereas they are constant under QS and $CRPS$.

The above results compare the score of the pool to the *average* score of its members. Another relevant comparison is between the pool and the best and worst of its members.

Corollary 3.2. *1.* $LS(y, \sum_{i=1}^{n} \omega_i f_i(\cdot)) \geq min_{i \in \{1,\ldots,n\}} LS(y, f_i(\cdot))$, *and analogously for* QS *and* $CRPS$.

2. $max_{i \in \{1,\ldots,n\}} LS(y, f_i(\cdot)) \geq LS(y, \sum_{i=1}^{n} \omega_i f_i(\cdot))$, *but analogous statements for* QS *and* $CRPS$ *do not hold.*

Hence the pool necessarily performs better than its worst member; this has been remarked by Kascha and Ravazzolo (2010) for the log score. Also, the log score of the pool is limited by the log score of its best member. Interestingly, the latter statement does not hold true for the other two scoring rules, where it is possible to construct examples in which the pool outperforms *all* of its members.

Figure 1 summarizes the results of Proposition 3.1, Corollary 3.1 and Corollary 3.2 for a simple example: An equally weighted combination of two Gaussian densities with mean zero and variances 1 and 4. The formulas which underlie the graphs are given in Table 1; they will also be used in the simulation study and the empirical application below. One of these formulas, the quadratic score for a pool of normals, has not been mentioned in the extant literature. A derivation of this formula is provided in the appendix.

It can be seen from the graphs that for all scoring rules and realizing outcomes, the score of the pool is above the lower bound and thus closer to the (ex post) better of the two models. For LS, the sharpness of the lower bound depends on the realizing outcome; this is not the case under QS and $CRPS$. Furthermore, the log score of the pool is bounded between the two log scores of its members. For the other two scoring rules, there is an (albeit small) range of realizing outcomes for which the pool outperforms both of its members.

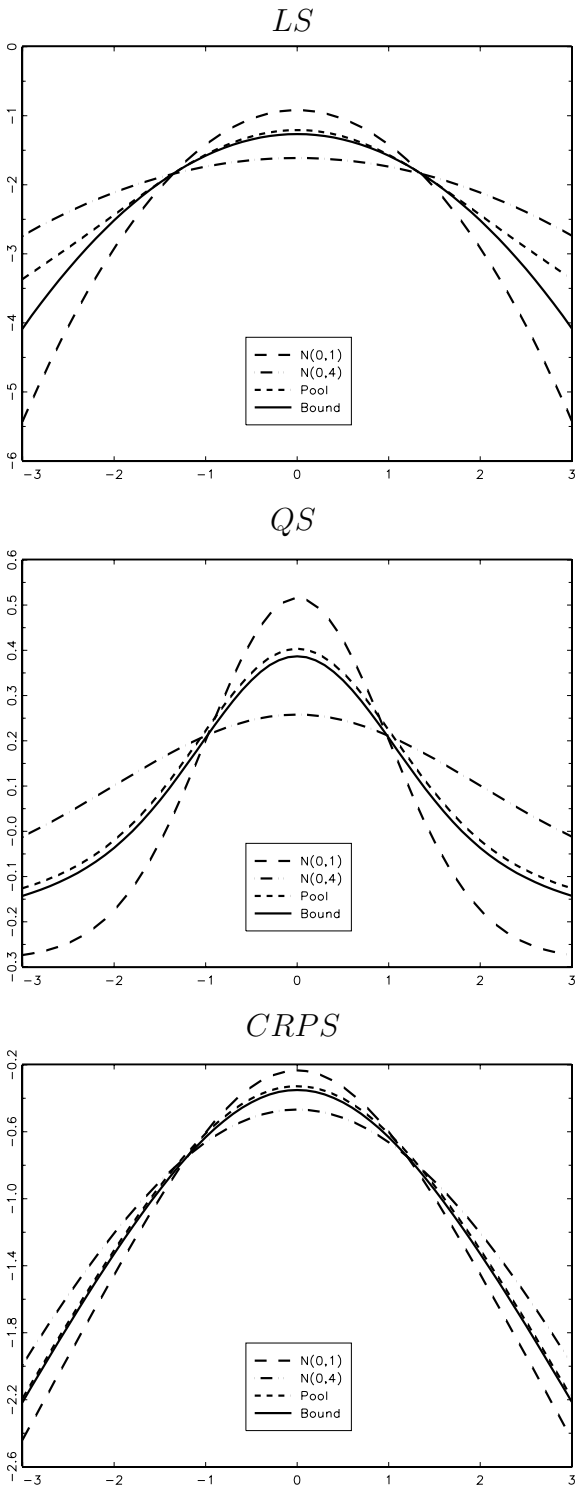

Figure 1: The upper graph plots the log score (vertical axis) against the realizing outcome (horizontal axis), for a $\mathcal{N}(0,1)$ density, a $\mathcal{N}(0,4)$ density, and an equally weighted linear pool. Furthermore, the graph shows the lower bound from Proposition 3.1. The graphs in the middle and the bottom give the same information for the quadratic and continuous ranked probability scores. The formulas which underlie these graphs are given in Table 1 below.

A simple but powerful consequence of Proposition 3.1 is that the *expected* score of a linear pool is at least as large as the average expected score of its components.

Corollary 3.3. *Let $f_0(Y)$ be the true density of Y, and denote by $ELS_0(f(\cdot))$ the expected log score of a predictive density $f(\cdot)$, with respect to the true density $f_0(\cdot)$. Then,*

$$ELS_0(\sum_{i=1}^{n} \omega_i f_i(\cdot)) \geq \sum_{i=1}^{n} \omega_i ELS_0(f_i(\cdot)),$$

and analogous relations hold for QS and CRPS.

Corollary 3.3 defines a lower bound for the expected performance of the pool, relative to the expected performance of the n components. It is intriguing that this bound holds true for all possible true distributions $f_0(\cdot)$. This implies that a forecaster does not need to know anything about $f_0(\cdot)$ to know that *in expectation*, pooling will perform reasonably well relative to the individual models.

The existence of lower bounds on their performance is an attractive feature of linear pools. An interesting question is whether similar lower bounds exist for other methods for combining predictive densities. I briefly consider two of these methods in the following.

Based on a set of n predictive densities and positive weights which sum to one, a logarithmic pool (Winkler, 1968) is given by

$$f_{log}(Y) = \frac{\prod_{i=1}^{n} f_i^{\omega_i}(Y)}{\int \prod_{i=1}^{n} f_i^{\omega_i}(z)dz}. \tag{12}$$

Genest and Zidek (1986) remark that in contrast to linear pools, logarithmic pools are typically less dispersed and unimodal.

	$\mathcal{N}(\mu, \sigma^2)$	$\sum_{i=1}^{n} \omega_i \mathcal{N}(\mu_i, \sigma_i^2)$
LS	$\ln \phi\left(\frac{y-\mu}{\sigma}\right) - \ln \sigma$	$\ln\left(\sum_{i=1}^{n} \frac{\omega_i}{\sigma_i} \phi\left(\frac{y-\mu_i}{\sigma_i}\right)\right)$
QS	$\frac{2}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right) - \frac{1}{\sqrt{4\pi\sigma^2}}$	$2\left(\sum_{i=1}^{n} \frac{\omega_i}{\sigma_i} \phi\left(\frac{y-\mu_i}{\sigma_i}\right)\right) - \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\omega_i \omega_j}{\sigma_j} (2\pi(1+b_{ij}^2))^{-\frac{1}{2}} \times \exp\left\{-\frac{a_{ij}^2}{2}\left(\frac{1}{1+b_{ij}^2}\right)\right\}$, with $a_{ij} = \frac{\mu_i - \mu_j}{\sigma_j}, b_{ij} = \frac{\sigma_i}{\sigma_j}$
$CRPS$	$-\sigma\left(\frac{y-\mu}{\sigma}\left[2\Phi\left(\frac{y-\mu}{\sigma}\right) - 1\right] + 2\phi\left(\frac{y-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}}\right)$	$-\sum_{i=1}^{n} \omega_i A(y-\mu_i, \sigma_i^2) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i \omega_j A(\mu_i - \mu_j, \sigma_i^2 + \sigma_j^2)$, with $A(\mu, \sigma^2) = 2\sigma\phi\left(\frac{\mu}{\sigma}\right) + \mu\left(2\Phi\left(\frac{\mu}{\sigma}\right) - 1\right)$

Table 1: Log score (LS), quadratic score (QS) and continuous ranked probability score ($CRPS$) for a normal distribution (left column) and a mixture of n normals (right column), given that an outcome $y \in \mathbb{R}$ materializes. The formulas for the $CRPS$ have been derived by Gneiting, Raftery, Westveld, and Goldman (2005) and Grimit, Gneiting, Berrocal, and Johnson (2006). The formula for the QS of a pool of normals is derived in the appendix.

A beta-transformed linear pool (Gneiting and Ranjan, 2011) is given by

$$f_b(Y) = (\sum_{i=1}^{n} \omega_i f_i(Y)) \times b(\sum_{i=1}^{n} \omega_i F_i(Y)). \tag{13}$$

Here $b(z) = B(\alpha, \beta)^{-1} z^{\alpha-1}(1-z)^{\beta-1}$ is the p.d.f of the beta distribution with strictly positive parameters α and β, and $B(\cdot, \cdot)$ denotes the beta function. Based on their key result that a linear pool of calibrated components is necessarily uncalibrated, Gneiting and Ranjan (2011) propose the combination in (13) to restore calibration. It has two tuning parameters, α and β, which can either be fixed or fitted to training data.

The next proposition shows that Jensen's inequality does not necessarily apply to logarithmic and beta-transformed pools. This contrasts the situation for linear pools, where Jensen's inequality holds under all three scoring rules.

Proposition 3.2. *For the logarithmic pool defined in Equation (12), it holds that*

$$LS(y, f_{log}(\cdot)) \geq \sum_{i=1}^{n} \omega_i LS(y, f_i(\cdot)),$$

but similar inequalities for QS and CRPS do not hold. Furthermore, for the beta-transformed linear pool defined in Equation (13), Jensen's inequality does not hold under LS.

4 Simulation Evidence

The results discussed until now are for a stylized setting in which forecasts are made for a single realization of the random variable Y. In practice, however, forecasts are typically made for a sequence of T realizations $\{y_t\}_{t=1}^{T}$ of a time series process Y_t. The forecasts of n different models are made h steps ahead and are based on a sigma algebra

$\mathcal{F}_{t-h}, h > 0$ generated by information up to time $t-h$. Let $f_i(Y_t|\mathcal{F}_{t-h}) \equiv f_{i,t-h}(Y_t)$ be the i-th forecasting density, with $i = 1, \ldots, n$. In this setting, a linear prediction pool with weights $\omega_{i,t-h} \in \mathcal{F}_{t-h}$ is given by $f_{c,t-h}(Y_t) = \sum_{i=1}^{n} \omega_{i,t-h} f_{i,t-h}(Y_t)$.

Hence the performance of linear pools depends on i) the true process generating Y_t, ii) the set of forecasting densities to be combined, and iii) the combination weights. In the following, I analyze the impact of i) and ii) within a simulation study. For brevity, I abstract from the third aspect by assuming equal weights. Equal weights are of high practical interest due to their trivial implementation without estimation uncertainty. Furthermore, the theoretical results from the last section are clearest for equal weights. I report additional simulation results for other weighting schemes in an online appendix to this paper.

I consider a setting in which a forecast of a time series process Y_t is required at horizon $h = 1$. The n individual models are Gaussian first-order autoregressive (AR(1)) processes. In addition, I consider their equally weighted (EW) combination:

$$\begin{aligned} f_i(Y_t|\mathcal{F}_{t-1}) &= \mathcal{N}(\nu_i + \alpha_i Y_{t-1}, \sigma_i^2), \ i = 1, \ldots, n, \\ f_{EW}(Y_t|\mathcal{F}_{t-1}) &= \frac{1}{n} \sum_{i=1}^{n} f_i(Y_t|\mathcal{F}_{t-1}), \end{aligned} \tag{14}$$

with $\alpha_i < 1 \ \forall \ i$. Assume that the true process is given by model $i^* \in \{1, \ldots, n\}$, i.e. $f_0(Y_t|\mathcal{F}_{t-1}) = f_{i^*}(Y_t|\mathcal{F}_{t-1})$. Then, under any of the three scoring rules S,

$$\frac{1}{n} \sum_{i=1}^{n} ES_0(f_{i,t-1}(\cdot)|\mathcal{F}_{t-1}) \ \leq \ ES_0(f_{EW,t-1}(\cdot)|\mathcal{F}_{t-1}) < ES_0(f_{i^*,t-1}(\cdot)|\mathcal{F}_{t-1}), \tag{15}$$

where $ES_0(f_{i,t-1}(\cdot)|\mathcal{F}_{t-1})$ denotes the expected score of model i under the true density, conditional on $\mathcal{F}_{t-1}$. Note that the first inequality follows from Corollary 3.3, and the second (strict) inequality follows from strict propriety of the three scoring rules. Assuming that the process generating $\mathcal{F}_{t-1}$ is stationary, the law of iterated expectations implies

that the inequalities hold also unconditionally:

$$\frac{1}{n}\sum_{i=1}^{n} ES_0(f_{i,t-1}(\cdot)) \quad \leq \quad ES_0(f_{EW,t-1}(\cdot)) < ES_0(f_{i^*,t-1}(\cdot)). \tag{16}$$

The second inequality in (16) reflects the disutility from using an equally weighted pool, instead of the true model i^*. In order to analyze the empirical relevance of this disutility, I consider the test statistic analyzed by Giacomini and White (2006):

$$GW_{i^*,EW} = \frac{\frac{1}{T}\sum_{t=1}^{T}\left(S(y_t, f_{i^*,t-1}(\cdot)) - S(y_t, f_{EW,t-1}(\cdot))\right)}{\sqrt{\hat{\sigma}^2_{i^*,EW}}}, \tag{17}$$

where $t = 1, \ldots, T$ denotes the evaluation sample. Note that the numerator of (17) is the difference between the average scores of the two models i^* and EW, and $\hat{\sigma}^2_{i^*,EW}$ is a Newey and West (1987, HAC) estimator of the variance of this difference. The statistic can be used to test the null hypothesis of equal predictive ability (EPA):

$$H_0 : ES_0(f_{i^*,t-1}(\cdot)) = ES_0(f_{EW,t-1}(\cdot)). \tag{18}$$

Under this null hypothesis and regularity conditions detailed in Giacomini and White (2006), the test statistic in (17) has a limiting standard normal distribution.

Of course, Equation (16) implies that the null hypothesis of EPA is false and should be rejected in favor of the true model i^*. Analyzing the test's actual rejection frequency (i.e., its power) allows to assess how often an empirical researcher would be able to statistically distinguish between the equally weighted pool and the true model. In the following, I report how often the null hypothesis is rejected in favor of i^* at a 5 % level, i.e. how often $GW_{i^*,EW} > 1.96$ in (17).

The processes in the following simulation experiment are calibrated to match four monthly US macro series: The consumer price index (CPI), industrial production, the treasury bill

rate, and the unemployment rate. All data were downloaded from the FRED® database of the Federal Reserve Bank at St. Louis and are transformed to achieve stationarity. For details, see Table 2 and Figure 3 in the appendix. For a given series, the model suite consists of $n = 5$ AR(1) models, each of which is calibrated to a different sample of empirical data (starting in 1960, 1970, 1980, 1990, and 2000; all samples end in November 2011). Table 3 in the appendix reports the resulting parameter values.

Tables 4 and 5 in the appendix report the simulation results, which are qualitatively very similar across the three scoring rules. When the true process is calibrated to industrial production or unemployment, the Giacomini and White (2006) has very little power in distinguishing between the equally weighted pool and the true model. Rejection rates for the small (large) evaluation sample average merely 9 (14) percent. The power of the tests is somewhat larger when the true process is calibrated to match CPI inflation. Still, rejection rates are only 18 (38) percent on average. In contrast, the EPA tests have good power properties when the true process is calibrated to the treasury bill rate, with average rejection rates of 69 (94) percent.

These results can be explained by the set of models used in the simulation experiments. As detailed in Table 3, the model coefficients for industrial production and unemployment are relatively stable over different subperiods. Hence for these series, the five models to be combined are fairly similar, which makes it hard for the EPA tests to distinguish between the equally weighted pool and any individual model. The opposite scenario applies to the treasury bill rate: The process features strong structural breaks in both the autoregressive coefficient and the error term variance. As a consequence, it is fairly easy to distinguish the equally weighted pool from any of the individual models. The inflation rate is somewhere between these two polar cases, featuring instabilities in AR persistence but not in the other two coefficients.[1]

[1]See Stock and Watson (2007) and Guidolin and Timmermann (2009) for additional evidence on instabilities in US inflation and short term interest rates, respectively.

I view these results as very positive for linear pools, especially because the setup of the simulation study is deliberately biased in favor of the individual models: The true process is given by one of these models, which of course need not hold in practice. In addition, I use a pool with equal weights. In contrast, Geweke and Amisano (2011) consider estimating unconditionally optimal weights based on a historical training sample. Bache, Jore, Mitchell, and Vahey (2011, Equation 2), among others, consider recursive weights based on the individual models' past performance. Both schemes can be expected to push the weights toward their theoretical optimum which has $\omega_{i^*,t-1} = 1$ and $\omega_{i,t-1} = 0$ for $i \neq i^*$. The results for equally weighted pools should thus give a conservative estimate of the performance of more general pools. The results for the other weighting schemes, which are presented in the online appendix, confirm this suggestion.

5 Empirical Evidence

The preceding simulation study is driven by the conservative assumption that the true process is given by one of the n individual models. I next drop this assumption and analyze the performance of linear pools in an empirical application where the true process is unknown.

5.1 Setup and Data

I again consider the four US macro series introduced in the last section. I use data between January 1985 and November 2011 (=323 observations) as an evaluation period; earlier observations are used for estimating the models.

The individual predictive models are several Gaussian (vector) autoregressive specifications (VARs). I consider iterated forecasts (Marcellino, Stock, and Watson, 2006) at horizons of one, three, and six months. The autoregressive models to be combined differ

along two dimensions: The set of system variables, and the sampling scheme used for parameter estimation. The set of system variables can in principle be validated using information criteria. In contrast, in the likely presence of structural breaks of unknown size and/or timing there is no clear-cut way of choosing the optimal estimation sample. Hence averaging over several estimation samples is a natural idea which has been proposed by Pesaran and Timmermann (2007) for point forecasts; Jore, Mitchell, and Vahey (2010) extend this idea to density forecasts.

Here I consider two different choices for the estimation sample: A short rolling window covering seven years of monthly data, and a long rolling window covering fourteen years. Furthermore, for each variable I consider four different sets of system variables: A univariate specification, and three bivariate specifications, each of which includes one of the other variables in addition. This leads to $2 \times 4 = 8$ forecast models for each variable. For all models, the lag lengths are adaptively chosen via the Schwarz (1978) information criterion using a maximum lag length of six. For simplicity, all VARs feature multivariate Gaussian, conditionally homoscedastic forecast densities.[2]

Throughout, I estimate all models via Ordinary Least Squares (OLS). A more stringent approach would be to tailor the parameter estimates to the scoring rule which is used for out-of-sample evaluation (c.f. Gneiting, 2011).[3] This is cumbersome for the $CRPS$, where multivariate generalizations (Gneiting, Stanberry, Grimit, Held, and Johnson, 2008), which are needed to estimate the VARs, are not available in closed form. Since these issues are out of the focus of the present paper, I follow standard practice and use OLS estimation instead.

[2]Rolling window estimation and adaptive lag length choice produce time variation in the estimated variance-covariance matrix of a VAR's error term vector. However, these effects are distinct from GARCH-type models in the tradition of Engle (1982) which would feature conditional heteroscedasticity even if one knew the true values of their parameters.

[3]Using OLS estimation, this principle is satisfied for the log score: Since the models are Gaussian, and the OLS and Maximum Likelihood estimators of a Gaussian model coincide, the parameter estimates maximize the in-sample log score of each model. However, the OLS estimates do not maximize the in-sample quadratic and continuous ranked probability scores.

As in the last section, I focus on equally weighted pools for brevity. Results for two other weighting schemes are presented in the online appendix.

5.2 Forecast Evaluation

The present empirical analysis covers four variables, three forecast horizons and three scoring rules, which amounts to 36 forecast horse races. Each horse race features nine competing models (eight autoregressive models and their equally weighted combination, EW). The results are summarized in Tables 6 to 9 in the appendix. The main findings are as follows.

First, for inflation and the treasury bill rate there is strong evidence that short rolling windows are preferable to long rolling windows. This finding, which holds across all forecast horizons and scoring rules, is in line with well known instabilities in the two series (see Section 4). These instabilities work in the favor of short rolling windows which discount outdated information more quickly than long windows. For industrial production and unemployment, the relative performance of short and long rolling windows is less clear and differs across system variables, forecast horizons and scoring rules. Quite generally, models using different system variables (but the same choice of estimation sample) tend to perform very similarly.

Second, the overall performance of EW is very good. Considering the average score over the evaluation sample, EW beats *all* eight individual models in half of the 36 horse races. Furthermore, the rank of EW among all nine models is never worse than five. The example in Figure 2, which refers to one-step ahead inflation forecasts under LS, provides some intuition for this result. The figure plots the ranks of EW and a $\text{VAR}^s_{C,T}$ model, for each point in the evaluation sample.[4] The performance of the $\text{VAR}^s_{C,T}$ (relative to

[4]The $\text{VAR}^s_{C,T}$ contains the CPI and treasury bill series and is estimated based on a short rolling window; see below Table 6. For each evaluation period t, the rank of a model is computed as one plus the number of models which attain a strictly higher score in period t.

the other models) is very instable over time. During the evaluation sample, it repeatedly visits all possible ranks (one to nine); this is illustrated by the wildly fluctuating light line. In contrast, by Jensen's inequality EW performs relatively well *for each point in the evaluation sample*. In the graph, this is reflected by the stable bold line. This stability is rewarded by concave scoring rules like the log score, so that EW outperforms the $\text{VAR}^s_{C,T}$ on average over the evaluation sample. Hence the example, and the empirical results in general, provide clear evidence against the classic notion of a "single best model".

Third, in order to analyze the statistical significance of these results, I consider Giacomini and White (2006) tests of equal predictive ability (EPA) as described in Section 4. The null hypothesis is EPA of the pool and a given individual model.[5] The results of the EPA tests are reported in Tables 6 to 9; see below Table 6 for details. The reported p-values refer to a two-sided test, and a truncation lag of four is used for the HAC estimator. The results feature an asymmetric pattern: While there are many cases in which EW significantly outperforms an individual model at the 5% level, it rarely happens that EW is itself outperformed. Even in horse races in which EW ranks fifth, it occurs that four models are significantly worse than EW while no model is significantly better. This is what happens for inflation at horizons three and six, using QS or $CRPS$. The only exception in which EW is repeatedly outperformed in terms of EPA tests is for the treasury bill rate and evaluation via QS. However, these results are not shared by LS and $CRPS$.

Fourth, there is evidence that the payoff to linear pooling is larger under LS than under QS and $CRPS$. Under LS, it *never* occurs (in $8 \times 4 \times 3 = 96$ occasions) that an individual model is significantly better than EW. Among others, EW performs well for inflation and the treasury bill rate. This is somewhat surprising since for these series, half of the models (the ones using long rolling windows) are clearly dominated by the other half. One might

[5]Note that unlike the classical test of Diebold and Mariano (1995), the test of Giacomini and White (2006) compares the predictive accuracy of the two *estimated* models, rather than the accuracy of their population counterparts. See Section 3 in Clark and McCracken (2011) for an insightful discussion.

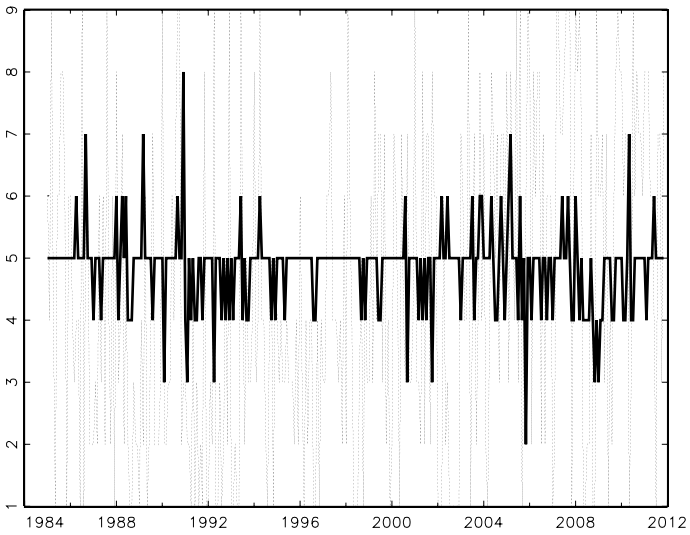

Figure 2: Light and bold lines: Rank of $\text{VAR}^s_{C,T}$ and EW, for each evaluation point, in terms of the log score for one-step ahead inflation forecasts, plotted against time.

expect that in such a setting, an equally weighted combination of *all* models would be inferior to the (ex post) better set of models. In fact, this is what happens under QS and $CRPS$, but not under LS. Hence under LS, equally weighted pools seem to be very robust to the inclusion of models which (ex post) turn out to perform poorly.

6 Conclusion

The results of this paper imply that linear pooling is an attractive forecasting strategy under three strictly proper scoring rules. The downside risk of linear pooling is bounded by Jensen's inequality. The existence of this bound is a mechanical consequence of the scoring rules and does not depend on the true data-generating process or the set of forecasting models being used. In contrast, the upside potential of linear pools does depend on these factors. I therefore present simulation and empirical evidence that equally weighted pools often perform similar to (or better than) the best individual model. This is remarkable since picking the best individual model requires hindsight, while constructing the equally weighted pool does not.

I next relate my results to recent studies which point to some negative features of linear prediction pools. Generalizing earlier results by Hora (2004) and Ranjan and Gneiting (2010), Gneiting and Ranjan (2011) show that a linear pool of well-calibrated densities will generally be overdispersed,[6] which can be diagnosed via Probability Integral Transforms (PITs). Clearly, this feature is a drawback of linear pools. However, these results do not contradict the findings of the present paper: It is well possible that an overdispersed density (in the definition of Gneiting and Ranjan) outperforms neutrally dispersed competitors in terms of scoring rules. The simulation example in Section 4.1 of Gneiting and Ranjan (2011) provides clear evidence in favor of this claim. In this example, the (overdispersed) linear pool outperforms all three neutrally dispersed competitors in terms of the average log score over the evaluation sample. See Tables 6 and 7 of Gneiting and Ranjan (2011). Hence this example leads to a situation in which different evaluation methodologies – PITs versus the log score – lead to different judgments about linear prediction pools. Although this situation is unsatisfactory from a practical perspective, its occurrence is not surprising in a setup with misspecification.

Interestingly, the tendency of linear pools to produce dispersed predictive densities is both a blessing and a curse. It can explain both the overdispersion results cited above *and* the lower bounds derived in the present paper. Compared to nonlinear combination methods, or the selection of a single model, linear pools are a safe choice which may entail some costs in terms of efficiency. Whether these costs are acceptable depends on the empirical scenario at hand; see also the comments by Gneiting and Ranjan (2011, p. 33). In macroeconomics, time series data is typically scarce and prone to structural breaks. Furthermore, crucial tuning parameters like the choice of estimation sample are hard to validate. In this scenario, the quest for (large-sample) optimality may be unrealistic, and a robust but suboptimal choice like equally weighted linear pools may do the job.

[6]See Definition 2.7 of Gneiting and Ranjan (2011) for a formal definition of overdispersion.

References

Bache, I. W., A. S. Jore, J. Mitchell, and S. P. Vahey (2011): "Combining VAR and DSGE Forecast Densities," *Journal of Economic Dynamics and Control*, 35, 1659 – 1670.

Boero, G., J. Smith, and K. F. Wallis (2011): "Scoring Rules and Survey Density Forecasts," *International Journal of Forecasting*, 27, 379 – 393.

Brier, G. W. (1950): "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, 78, 1–3.

Clark, T. E., and M. W. McCracken (2011): "Advances in Forecast Evaluation," Working Paper, Federal Reserve Bank of St. Louis (May 2011).

Clements, M. P., and D. I. Harvey (2011): "Combining Probability Forecasts," *International Journal of Forecasting*, 27, 208 – 223.

Diebold, F. X., and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253–63.

Engle, R. F. (1982): "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50, 987–1007.

Genest, C., and J. V. Zidek (1986): "Combining Probability Distributions: A Critique and an Annotated Bibliography," *Statistical Science*, 1, 114–135.

Geweke, J. W., and G. Amisano (2011): "Optimal Prediction Pools," *Journal of Econometrics*, 164, 130–141.

Giacomini, R., and H. White (2006): "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545–1578.

Gneiting, T. (2011): "Making and Evaluating Point Forecasts," *Journal of the American Statistical Association*, 106, 746–762.

Gneiting, T., and A. E. Raftery (2007): "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman (2005): "Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation," *Monthly Weather Review*, 133, 1098–1118.

Gneiting, T., and R. Ranjan (2011): "Combining Predictive Distributions," Working Paper, University of Heidelberg (June 9, 2011).

Gneiting, T., and R. Ranjan (2011): "Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules," *Journal of Business and Economic Statistics*, 29, 411–422.

Gneiting, T., L. Stanberry, E. Grimit, L. Held, and N. Johnson (2008): "Assessing Probabilistic Forecasts of Multivariate Quantities, with an Application to Ensemble Predictions of Surface Winds," *TEST*, 17, 211–235.

Gneiting, T., and T. L. Thorarinsdottir (2010): "Predicting Inflation: Professional Experts Versus No-Change Forecasts," Working Paper, University of Heidelberg (October 13, 2010).

Good, I. (1952): "Rational Decisions," *Journal of the Royal Statistical Society, Series B*, 14, 107–114.

Grimit, E. P., T. Gneiting, V. J. Berrocal, and N. A. Johnson (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.

Guidolin, M., and A. Timmermann (2009): "Forecasts of US Short-Term Interest Rates: A Flexible Forecast Combination Approach," *Journal of Econometrics*, 150, 297 – 311.

Hall, S. G., and J. Mitchell (2007): "Combining Density Forecasts," *International Journal of Forecasting*, 23, 1–13.

Hersbach, H. (2000): "Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems," *Weather and Forecasting*, 15, 559–570.

Hora, S. C. (2004): "Probability Judgments for Continuous Quantities: Linear Combinations and Calibration," *Management Science*, 50, 597–604.

Jore, A. S., J. Mitchell, and S. P. Vahey (2010): "Combining Forecast Densities from VARs with Uncertain Instabilities," *Journal of Applied Econometrics*, 25, 621–634.

Kascha, C., and F. Ravazzolo (2010): "Combining Inflation Density Forecasts," *Journal of Forecasting*, 29, 231–250.

Kullback, S., and R. A. Leibler (1951): "On Information and Sufficiency," *The Annals of Mathematical Statistics*, 22, 79–86.

Manski, C. F. (2010): "When Consensus Choice Dominates Individualism: Jensen's Inequality and Collective Decisions Under Uncertainty," *Quantitative Economics*, 1, 187–202.

Marcellino, M., J. H. Stock, and M. W. Watson (2006): "A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series," *Journal of Econometrics*, 135, 499–526.

Matheson, J. E., and R. L. Winkler (1976): "Scoring Rules for Continuous Probability Distributions," *Management Science*, 22, 1087–1096.

McNees, S. K. (1992): "The Uses and Abuses of "Consensus" Forecasts," *Journal of Forecasting*, 11, 703–710.

Newey, W. K., and K. D. West (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703 – 708.

Pesaran, M. H., and A. Timmermann (2007): "Selection of Estimation Window in the Presence of Breaks," *Journal of Econometrics*, 137, 134 – 161.

Ranjan, R., and T. Gneiting (2010): "Combining Probability Forecasts," *Journal of the Royal Statistical Society, Series B*, 72, 71–91.

Schwarz, G. (1978): "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Selten, R. (1998): "Axiomatic Characterization of the Quadratic Scoring Rule," *Experimental Economics*, 1, 43–62.

Stock, J. H., and M. W. Watson (2007): "Why Has U.S. Inflation Become Harder to Forecast?," *Journal of Money, Credit and Banking*, 39, 3–33.

Stone, M. (1961): "The Opinion Pool," *Annals of Mathematical Statistics*, 32, 1339–1342.

Timmermann, A. (2006): "Forecast Combinations," vol. 1 of *Handbook of Economic Forecasting*, pp. 135 – 196. Elsevier.

Wallis, K. F. (2005): "Combining Density and Interval Forecasts: A Modest Proposal," *Oxford Bulletin of Economics and Statistics*, 67, 983–994.

Winkler, R. L. (1968): "The Consensus of Subjective Probability Distributions," *Management Science*, 15, B61–B75.

——— (1996): "Scoring Rules and the Evaluation of Probabilities," *TEST*, 5, 1–26.

Appendix A: Proofs

Proof of Proposition 3.1

$$\begin{aligned} LS(y, \sum_{i=1}^{n} \omega_i f_i(\cdot)) &= \ln\left(\sum_{i=1}^{n} \omega_i f_i(y)\right) \\ &\geq \sum_{i=1}^{n} \omega_i \ln\left(f_i(y)\right) \\ &= \sum_{i=1}^{n} \omega_i LS(y, f_i(\cdot)), \end{aligned}$$

where the inequality follows from Jensen's inequality.

$$\begin{aligned} QS(y, \sum_{i=1}^{n} \omega_i f_i(\cdot)) &= \sum_{i=1}^{n} 2\omega_i f_i(y) - \int \underbrace{(\sum_{i=1}^{n} \omega_i f_i(z))^2}_{\leq \sum_{i=1}^{n} \omega_i f_i^2(z)} dz \\ &\geq \sum_{i=1}^{n} 2\omega_i f_i(y) - \sum_{i=1}^{n} \omega_i \int f_i^2(z) dz \\ &= \sum_{i=1}^{n} \omega_i \left(2f_i(y) - \int f_i^2(z) dz\right) \\ &= \sum_{i=1}^{n} \omega_i QS(y, f_i(\cdot)), \end{aligned}$$

where the inequality follows from Jensen's inequality.

$$\begin{aligned} CRPS(y, \sum_{i=1}^{n} \omega_i f_i(\cdot)) &= -\int \left(\sum_{i=1}^{n} \omega_i F_i(z) - \mathbb{I}(z \geq y)\right)^2 dz \\ &= -\int \underbrace{\left(\sum_{i=1}^{n} \omega_i (F_i(z) - \mathbb{I}(z \geq y))\right)^2}_{\leq \sum_{i=1}^{n} \omega_i (F_i(z) - \mathbb{I}(z \geq y))^2} dz \\ &\geq -\sum_{i=1}^{n} \omega_i \int \left(F_i(z) - \mathbb{I}(z \geq y)\right)^2 dz \\ &= \sum_{i=1}^{n} \omega_i CRPS(y, f_i(\cdot)), \end{aligned}$$

where the second equality uses the fact that the weights sum to one and the inequality follows

from Jensen's inequality.

Proof of Corollary 3.1

$\Delta_{LS} = LS(y, \sum_{i=1}^n \omega_i f_i(\cdot)) - \sum_{i=1}^n \omega_i LS(y, f_i(\cdot))$, and similarly for QS and $CRPS$. From the definition of LS in Equation (1),

$$\begin{aligned}
\Delta_{LS} &= \ln(\sum_{i=1}^n \omega_i f_i(y)) - \underbrace{\sum_{i=1}^n \omega_i \ln(f_i(y))}_{=\sum_{i=1}^n \ln(f_i(y)^{\omega_i}) = \ln(\prod_{i=1}^n f_i(y)^{\omega_i})} \\
&= \ln\left(\frac{\sum_{i=1}^n \omega_i f_i(y)}{\prod_{i=1}^n f_i(y)^{\omega_i}}\right).
\end{aligned}$$

The formula for Δ_{QS} follows immediately from the definition of QS in Equation (2). Furthermore, from the definition of $CRPS$ in (3),

$$\begin{aligned}
\Delta_{CRPS} &= \int \left\{ \sum_{i=1}^n \omega_i (F_i(z) - \mathbb{I}_{\{z \geq y\}})^2 - (\sum_{i=1}^n \omega_i F_i(z) - \mathbb{I}_{\{z \geq y\}})^2 \right\} dz \\
&= \int \left\{ \sum_{i=1}^n \omega_i (F_i^2(z) - 2\, \mathbb{I}_{\{z \geq y\}} F_i(z) + \mathbb{I}_{\{z \geq y\}}) - (\sum_{i=1}^n \omega_i F_i(z))^2 + 2\, \mathbb{I}_{\{z \geq y\}} \sum_{i=1}^n \omega_i F_i(z) - \mathbb{I}_{\{z \geq y\}} \right\} dz \\
&= \int \left\{ \sum_{i=1}^n \omega_i F_i^2(z) - (\sum_{i=1}^n \omega_i F_i(z))^2 \right\} dz,
\end{aligned}$$

where the second equality uses the fact that a binary term equals its square and the third equality uses the fact that the weights sum to one.

The positivity of the three terms follows from Proposition 3.1 or directly from Jensen's inequality.

Proof of Example 3.1

For $i = 1, 2$, let $f_i(\cdot) = \mathcal{N}(\mu_i, \sigma_i^2)$. Without loss of generality, assume that $\sigma_1^2 > \sigma_2^2$. For a weight $\omega \in (0, 1)$, it holds that

$$\begin{aligned}
\Delta_{LS} &= \ln\left(\omega f_1(y) + (1-\omega) f_2(y)\right) - \omega \ln(f_1(y)) - (1-\omega)\ln(f_2(y)) \\
&\geq \ln(\omega f_1(y)) - \omega \ln(f_1(y)) - (1-\omega)\ln(f_2(y)) \\
&= \ln(\omega) + (1-\omega)\left(\ln(f_1(y)) - \ln(f_2(y))\right) \\
&= \ln(\omega) + (1-\omega)\left(0.5\ln(2\pi\sigma_2^2) + \frac{(y-\mu_2)^2}{2\sigma_2^2} - 0.5\ln(2\pi\sigma_1^2) - \frac{(y-\mu_1)^2}{2\sigma_1^2}\right) \\
&= \left(\underbrace{\frac{1-\omega}{2}}_{>0}\left(\underbrace{\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}}_{>0}\right)\right) y^2 + \left((1-\omega)\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right)\right) y + c, \qquad (19)
\end{aligned}$$

where c does not depend on y. The limiting behavior of (19) as $y \to \infty$ is governed by the first (quadratic) term. Since the coefficient in front of this term is strictly positive, (19) – and thus also Δ_{LS} – diverge to ∞ as $y \to \infty$. The same reasoning applies when $y \to -\infty$.

Proof of Corollary 3.2

The first part follows from Proposition 3.1 and the fact that $\sum_{i=1}^{n} \omega_i LS(y, f_i(\cdot)) \geq \min_{i \in \{1,\dots,n\}} LS(y, f_i(\cdot))$, which in turn follows from the positivity of the weights. Analogous reasoning applies for QS and $CRPS$.

Concerning the second part,

$$\begin{aligned} LS(y, \sum_{i=1}^{n} \omega_i f_i(\cdot)) &= \ln(\sum_{i=1}^{n} \omega_i f_i(y)) \\ &\leq \ln(\max_{i \in \{1,\dots,n\}} f_i(y)) \\ &= \max_{i \in \{1,\dots,n\}} LS(y, f_i(\cdot)). \end{aligned}$$

In contrast to this result, Figure 1 provides examples which show that under QS and $CRPS$, a linear pool can potentially outperform all of its components. Among other cases, this happens under QS when $y = 1$, and under $CRPS$ when $y = 1.1$.

Proof of Corollary 3.3

$$\begin{aligned} ELS_0\left(\sum_{i=1}^{n} \omega_i f_i(\cdot)\right) &= \int \underbrace{LS(z, \sum_{i=1}^{n} \omega_i f_i(\cdot))}_{\geq \sum_{i=1}^{n} \omega_i LS(z, f_i(\cdot))} f_0(z) dz \\ &\geq \sum_{i=1}^{n} \omega_i \int LS(z, f_i(\cdot)) f_0(z) dz \\ &= \sum_{i=1}^{n} \omega_i ELS_0\left(f_i(\cdot)\right), \end{aligned}$$

where the inequality follows from Proposition 3.1. The proof for the quadratic and continuous ranked probability scores is analogous.

Proof of Proposition 3.2

The proof that Jensen's inequality holds for the logarithmic pool under the log score follows along the lines of Kascha and Ravazzolo (2010, p. 237), which can easily be extended to $n > 2$ models.

To show that Jensen's inequality does not hold for QS and $CRPS$, it is enough to find coun-

terexamples which violate the inequality. To get such examples, consider again the setup of Figure 1, where the first model is $f_1(\cdot) = \mathcal{N}(0, 1)$, the second model is $f_2(\cdot) = \mathcal{N}(0, 4)$, and the combination weights are 0.5. In this case, the logarithmic pool is given by $f_{\log}(\cdot) = \mathcal{N}(0, \frac{8}{5})$ (c.f. Kascha and Ravazzolo, 2010, p. 235). Suppose that the outcome $y = 2.5$ realizes. Using the formulas in Table 1 gives

$$\begin{aligned} \underbrace{QS(2.5, f_{log}(\cdot))}_{\approx -0.13} &< 0.5 \times \underbrace{QS(2.5, f_1(\cdot))}_{\approx -0.25} + 0.5 \times \underbrace{QS(2.5, f_2(\cdot))}_{\approx 0.04}, \\ \underbrace{CRPS(2.5, f_{log}(\cdot))}_{\approx -1.81} &< 0.5 \times \underbrace{CRPS(2.5, f_1(\cdot))}_{\approx -1.94} + 0.5 \times \underbrace{CRPS(2.5, f_2(\cdot))}_{\approx -1.57}, \end{aligned}$$

which contradicts Jensen's inequality.

To find an example in which Jensen's inequality does not hold for the beta-transformed linear pool under the log score, consider again the two individual densities and weights from above. Furthermore, for the beta-transformed linear pool $f_b(\cdot)$, set $\alpha = 1.492$ and $\beta = 1.440$. These numbers match the empirical results of Gneiting and Ranjan (2011, Table 5). Again considering a realization of $y = 2.5$ and using the formulas of Gneiting and Ranjan (2011, p. 21), one finds

$$\underbrace{LS(2.5, f_b(\cdot))}_{\approx -3.33} < 0.5 \times \underbrace{LS(2.5, f_1(\cdot))}_{\approx -4.04} + 0.5 \times \underbrace{LS(2.5, f_2(\cdot))}_{\approx -2.39},$$

which again contradicts Jensen's inequality.

Quadratic Score of a Gaussian Mixture

Auxiliary Result

Let a and b be two real-valued constants, and denote by $\phi(\cdot)$ the p.d.f. of the standard normal distribution. Then,

$$\begin{aligned}
\int \phi(u)\phi(a+bu)\,du &= \frac{1}{2\pi}\int \exp\left(-\frac{u^2}{2}-\frac{(a+bu)^2}{2}\right)du \\
&= \frac{1}{2\pi}\int \exp\left(-\frac{(1+b^2)u^2}{2}-\frac{2abu}{2}\right)\exp\left(-\frac{a^2}{2}\right)du \\
&= \frac{1}{2\pi}\exp\left(-\frac{a^2}{2}+\frac{a^2b^2}{2(1+b^2)}\right)\int \exp\left(-\frac{(1+b^2)u^2}{2}-\frac{2abu}{2}-\frac{a^2b^2}{2(1+b^2)}\right)du \\
&= \frac{1}{2\pi}\exp\left(-\frac{a^2}{2}+\frac{a^2b^2}{2(1+b^2)}\right)\int \exp\left(-\frac{1}{2}\underbrace{\left(\frac{ab}{\sqrt{1+b^2}}+\sqrt{1+b^2}u\right)^2}_{=\frac{\left(\frac{ab}{1+b^2}+u\right)^2}{\left(\frac{1}{1+b^2}\right)}}\right)du \\
&= \frac{1}{2\pi}\exp\left(-\frac{a^2}{2}+\frac{a^2b^2}{2(1+b^2)}\right)\sqrt{\frac{2\pi}{1+b^2}}\underbrace{\int \frac{1}{\sqrt{\frac{2\pi}{1+b^2}}}\exp\left(-\frac{1}{2}\frac{\left(u-\left(-\frac{ab}{1+b^2}\right)\right)^2}{\frac{1}{1+b^2}}\right)du}_{=\text{ p.d.f. of }\mathcal{N}\left(-\frac{ab}{1+b^2},\frac{1}{1+b^2}\right)} \\
&= \left(2\pi(1+b^2)\right)^{-\frac{1}{2}}\exp\left(-\frac{a^2}{2}\left(1-\frac{b^2}{1+b^2}\right)\right)
\end{aligned}$$

Main Result

The quadratic score for a generic p.d.f. $f(\cdot)$ and a realization y is given by

$$QS\left(f(\cdot),y\right)=2f(y)-\int f^2(z)dz \tag{20}$$

If $f(\cdot)$ is a mixture of n normals,

$$f(y)=\sum_{i=1}^{n}\frac{\omega_i}{\sigma_i}\phi\left(\frac{y-\mu_i}{\sigma_i}\right) \tag{21}$$

Furthermore,

$$\begin{aligned}\int f^2(z)dz &= \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\omega_i\omega_j}{\sigma_i\sigma_j}\int \phi\left(\frac{z-\mu_i}{\sigma_i}\right)\phi\left(\frac{z-\mu_j}{\sigma_j}\right)dz \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\omega_i\omega_j}{\sigma_j}(2\pi(1+b_{ij}^2))^{-\frac{1}{2}}\times\exp\left\{-\frac{a_{ij}^2}{2}\left(\frac{1}{1+b_{ij}^2}\right)\right\},\end{aligned} \tag{22}$$

where a_{ij} and b_{ij} have been defined in Table 1, and the second equality follows from integration by substitution and the auxiliary result above. Combining (20), (21) and (22) then yields the result in Table 1.

Appendix B: Additional Figures and Tables

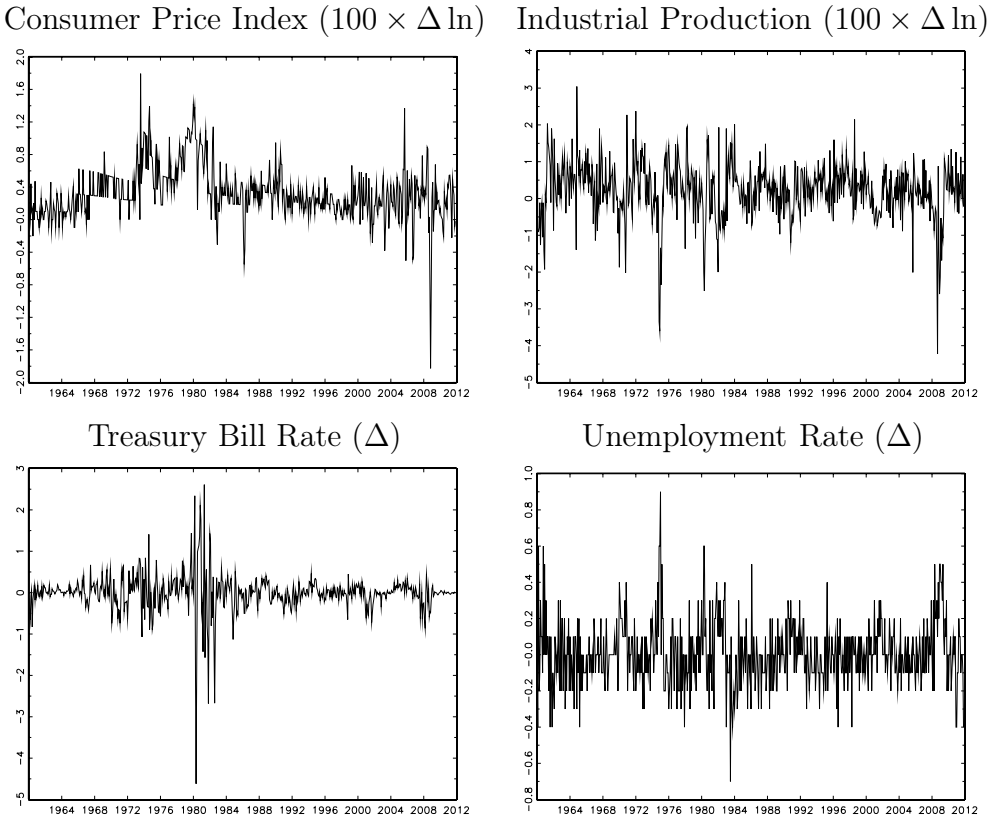

Figure 3: Time series graphs for the full sample (February 1960 – November 2011). See below Table 2 for details.

Variable	Description	Transformation	Mean	Std.	Q05	Q50	Q95	ACF(1)	ACF(5)
Consumer Price Index	All urban customers, seasonally adjusted, FRED series "CPIAUCSL"	$100 \times \Delta \ln$	0.33	0.32	-0.09	0.29	0.95	0.62	0.41
Industrial Production	Seasonally adjusted, FRED series "INDPRO"	$100 \times \Delta \ln$	0.22	0.78	-0.97	0.27	1.36	0.34	0.1
Treasury Bill Rate	Secondary market rate, FRED series "TB3MS"	Δ	-0.01	0.45	-0.62	0.01	0.51	0.33	0.04
Unemployment Rate	Percent, seasonally adjusted, FRED series "UNRATE"	Δ	0.01	0.18	-0.3	0	0.3	0.11	0.16

Table 2: Definitions and descriptive statistics. Transformation "$100 \times \Delta \ln$" means that $Y_t = 100 \times (\ln X_t - \ln X_{t-1})$, where Y_t is the transformed and X_t is the original series at date t. Similarly, transformation "Δ" means that $Y_t = X_t - X_{t-1}$. All descriptive statistics are for the transformed series over the full sample (February 1960 – November 2011, 623 monthly observations). "Q05", "Q50" and "Q95" denote the five, 50 and 95 percent quantiles. "ACF(1)" and "ACF(5)" denote the autocorrelation at lags one and five.

	$i = 1$ (1960-2011)	$i = 2$ (1970-2011)	$i = 3$ (1980-2011)	$i = 4$ (1990-2011)	$i = 5$ (2000-2011)
Consumer Price Index	0.124 0.624 0.062	0.131 0.631 0.068	0.122 0.554 0.062	0.128 0.412 0.061	0.116 0.431 0.096
Industrial Production	0.145 0.336 0.528	0.118 0.362 0.500	0.113 0.285 0.456	0.125 0.246 0.426	0.020 0.280 0.520
Treasury Bill Rate	-0.005 0.333 0.183	-0.010 0.332 0.218	-0.020 0.354 0.229	-0.016 0.458 0.032	-0.018 0.524 0.034
Unemployment Rate	0.005 0.116 0.033	0.007 0.197 0.032	0.005 0.173 0.030	0.010 0.166 0.025	0.022 0.296 0.027

Table 3: Parameter calibrations used in the simulation study. The numbers in each cell represent the parameters ν_i (first row), α_i (second row) and σ_i^2 (third row) of the AR(1) model in Equation (14), for four different time series. The series are transformed to stationarity; see Table 2 and Figure 3 for details.

		Consumer Price Index					*Industrial Production*				
	True model (i^*)	1	2	3	4	5	1	2	3	4	5
	$\frac{1}{n}\sum_{i=1}^{n} ES_0(f_{i,t-1}(\cdot))$	-0.066	-0.118	-0.047	-0.052	-0.314	-1.110	-1.080	-1.032	-1.002	-1.105
	$ES_0(f_{EW,t-1}(\cdot))$	-0.048	-0.096	-0.032	-0.038	-0.283	-1.104	-1.075	-1.027	-0.998	-1.100
LS	$ES_0(f_{i^*,t-1}(\cdot))$	-0.031	-0.076	-0.027	-0.022	-0.247	-1.100	-1.072	-1.026	-0.992	-1.092
	Rej. ($T = 120$)	0.208	0.174	0.191	0.328	0.263	0.043	0.033	0.117	0.186	0.118
	Rej. ($T = 360$)	0.440	0.434	0.267	0.551	0.679	0.106	0.074	0.123	0.264	0.232
	$\frac{1}{n}\sum_{i=1}^{n} ES_0(f_{i,t-1}(\cdot))$	1.094	1.040	1.114	1.109	0.862	0.384	0.396	0.416	0.429	0.386
	$ES_0(f_{EW,t-1}(\cdot))$	1.114	1.062	1.131	1.123	0.879	0.386	0.398	0.417	0.430	0.388
QS	$ES_0(f_{i^*,t-1}(\cdot))$	1.130	1.082	1.135	1.140	0.910	0.388	0.399	0.418	0.433	0.391
	Rej. ($T = 120$)	0.104	0.109	0.081	0.213	0.215	0.049	0.038	0.066	0.102	0.100
	Rej. ($T = 360$)	0.252	0.283	0.116	0.359	0.500	0.092	0.070	0.066	0.150	0.173
	$\frac{1}{n}\sum_{i=1}^{n} ES_0(f_{i,t-1}(\cdot))$	-0.145	-0.153	-0.142	-0.143	-0.180	-0.413	-0.401	-0.382	-0.370	-0.411
	$ES_0(f_{EW,t-1}(\cdot))$	-0.143	-0.150	-0.140	-0.142	-0.178	-0.411	-0.400	-0.381	-0.369	-0.410
$CRPS$	$ES_0(f_{i^*,t-1}(\cdot))$	-0.141	-0.147	-0.140	-0.140	-0.175	-0.410	-0.399	-0.381	-0.368	-0.407
	Rej. ($T = 120$)	0.118	0.141	0.088	0.262	0.252	0.043	0.038	0.098	0.145	0.132
	Rej. ($T = 360$)	0.304	0.374	0.135	0.465	0.519	0.091	0.080	0.097	0.195	0.230

Table 4: Simulation results. Horizontal blocks represent the log score (LS), quadratic score (QS) and cumulative ranked probability score ($CRPS$). In each block, the first three rows are simulation estimates of the quantities in Equation (16). All estimates are averages over 10000 Monte Carlo samples, each of which is 480 periods long. The fourth and fifth rows are rejection frequencies of the null hypothesis in (18), for two different sample sizes. The frequencies are computed over 10000 Monte Carlo samples. A truncation lag of four is used for the Newey and West (1987) estimator. Columns represent different true processes, calibrated to different subsamples of CPI inflation and industrial production. See Table 3 for details on calibration.

		Treasury Bill Rate					*Unemployment Rate*				
	True model (i^*)	1	2	3	4	5	1	2	3	4	5
	$\frac{1}{n}\sum_{i=1}^{n} ES_0(f_{i,t-1}(\cdot))$	-1.175	-1.445	-1.514	-0.010	-0.024	0.273	0.290	0.326	0.413	0.362
	$ES_0(f_{EW,t-1}(\cdot))$	-0.621	-0.736	-0.767	0.061	0.048	0.280	0.296	0.332	0.417	0.367
LS	$ES_0(f_{i^*,t-1}(\cdot))$	-0.570	-0.657	-0.681	0.299	0.275	0.286	0.298	0.332	0.424	0.378
	Rej. ($T = 120$)	0.436	0.594	0.634	0.998	0.997	0.064	0.039	0.058	0.184	0.149
	Rej. ($T = 360$)	0.860	0.962	0.975	1.000	1.000	0.157	0.070	0.064	0.279	0.303
	$\frac{1}{n}\sum_{i=1}^{n} ES_0(f_{i,t-1}(\cdot))$	0.450	0.368	0.349	1.232	1.211	1.535	1.561	1.617	1.766	1.676
	$ES_0(f_{EW,t-1}(\cdot))$	0.587	0.506	0.488	1.365	1.344	1.543	1.569	1.625	1.773	1.684
QS	$ES_0(f_{i^*,t-1}(\cdot))$	0.659	0.603	0.590	1.573	1.535	1.552	1.572	1.626	1.783	1.702
	Rej. ($T = 120$)	0.424	0.570	0.594	0.919	0.901	0.073	0.054	0.055	0.096	0.097
	Rej. ($T = 360$)	0.835	0.947	0.957	1.000	1.000	0.130	0.063	0.059	0.155	0.192
	$\frac{1}{n}\sum_{i=1}^{n} ES_0(f_{i,t-1}(\cdot))$	-0.254	-0.279	-0.286	-0.121	-0.124	-0.103	-0.102	-0.098	-0.090	-0.095
	$ES_0(f_{EW,t-1}(\cdot))$	-0.245	-0.270	-0.277	-0.113	-0.116	-0.103	-0.101	-0.098	-0.089	-0.095
$CRPS$	$ES_0(f_{i^*,t-1}(\cdot))$	-0.241	-0.264	-0.270	-0.101	-0.104	-0.103	-0.101	-0.098	-0.089	-0.094
	Rej. ($T = 120$)	0.306	0.460	0.541	0.999	0.998	0.080	0.034	0.069	0.142	0.099
	Rej. ($T = 360$)	0.711	0.905	0.945	1.000	1.000	0.151	0.055	0.072	0.202	0.243

Table 5: Simulation results (continued). True processes calibrated to the treasury bill rate and the unemployment rate. See Table 4 for details.

Consumer Price Index

Model	**LS**	**QS**	**CRPS**	**SE**
		$h = 1$		
AR^s	$0.02^{22.9}$	$1.49^{75.8}$	$-0.12^{63.3}$	$0.06^{76.2}$
AR^l	$-0.09^{2.5}$	$1.36^{0.0}$	$-0.13^{0.0}$	$0.06^{2.3}$
$\text{VAR}^s_{C,I}$	$0.00^{10.3}$	$1.46^{53.5}$	$-0.13^{7.8}$	$0.06^{4.7}$
$\text{VAR}^l_{C,I}$	$-0.11^{0.6}$	$1.33^{0.0}$	$-0.13^{0.0}$	$0.06^{0.7}$
$\text{VAR}^s_{C,T}$	$0.04^{39.1}$	$1.47^{78.7}$	$-0.13^{40.8}$	$0.06^{76.1}$
$\text{VAR}^l_{C,T}$	$-0.09^{1.0}$	$1.36^{0.0}$	$-0.13^{0.0}$	$0.06^{4.4}$
$\text{VAR}^s_{C,U}$	$-0.01^{6.7}$	$1.46^{41.5}$	$-0.13^{2.8}$	$0.06^{2.9}$
$\text{VAR}^l_{C,U}$	$-0.12^{0.8}$	$1.33^{0.0}$	$-0.13^{0.0}$	$0.06^{0.0}$
EW	0.06	1.48	−0.12	0.06
(Rank)	(1)	(2)	(2)	(2)
		$h = 3$		
AR^s	$-0.10^{58.3}$	$1.39^{41.7}$	$-0.13^{46.6}$	$0.08^{82.2}$
AR^l	$-0.27^{3.5}$	$1.20^{0.0}$	$-0.14^{0.0}$	$0.08^{0.5}$
$\text{VAR}^s_{C,I}$	$-0.09^{77.6}$	$1.37^{70.4}$	$-0.14^{75.3}$	$0.08^{40.3}$
$\text{VAR}^l_{C,I}$	$-0.28^{2.1}$	$1.18^{0.0}$	$-0.15^{0.0}$	$0.08^{0.1}$
$\text{VAR}^s_{C,T}$	$-0.09^{68.9}$	$1.38^{64.3}$	$-0.14^{99.2}$	$0.08^{70.2}$
$\text{VAR}^l_{C,T}$	$-0.28^{2.1}$	$1.17^{0.0}$	$-0.15^{0.0}$	$0.08^{0.0}$
$\text{VAR}^s_{C,U}$	$-0.09^{68.0}$	$1.37^{74.8}$	$-0.14^{58.9}$	$0.08^{31.0}$
$\text{VAR}^l_{C,U}$	$-0.29^{1.6}$	$1.17^{0.0}$	$-0.15^{0.0}$	$0.08^{0.0}$
EW	−0.08	1.36	−0.14	0.08
(Rank)	(1)	(5)	(2)	(1)
		$h = 6$		
AR^s	$-0.12^{85.8}$	$1.34^{51.0}$	$-0.14^{45.6}$	$0.08^{82.2}$
AR^l	$-0.31^{2.4}$	$1.13^{0.0}$	$-0.15^{0.0}$	$0.08^{0.3}$
$\text{VAR}^s_{C,I}$	$-0.11^{98.4}$	$1.34^{53.6}$	$-0.14^{43.4}$	$0.08^{65.8}$
$\text{VAR}^l_{C,I}$	$-0.33^{1.4}$	$1.10^{0.0}$	$-0.15^{0.0}$	$0.08^{0.0}$
$\text{VAR}^s_{C,T}$	$-0.11^{96.6}$	$1.34^{58.2}$	$-0.14^{57.8}$	$0.08^{93.0}$
$\text{VAR}^l_{C,T}$	$-0.33^{1.2}$	$1.09^{0.0}$	$-0.15^{0.0}$	$0.08^{0.0}$
$\text{VAR}^s_{C,U}$	$-0.11^{99.7}$	$1.34^{47.9}$	$-0.14^{39.3}$	$0.08^{61.8}$
$\text{VAR}^l_{C,U}$	$-0.33^{1.4}$	$1.11^{0.0}$	$-0.15^{0.0}$	$0.08^{0.0}$
EW	−0.11	1.32	−0.14	0.08
(Rank)	(2)	(5)	(5)	(5)

Table 6: Performance of density forecasts of the US consumer price index, for an evaluation sample from January 1985 to November 2011 (323 monthly observations), for forecast horizons of $h = 1$, 3 and 6 months. In column one, superscript s indicates that a model was estimated based on a short rolling window (72 obs.); l indicates a long rolling window (288 obs.). For the VARs, subscripts indicate the system variables, where C = CPI inflation, I = industrial production, T = treasury bill rate, U = unemployment rate. "EW" is an equally weighted pool of all models. The second to third columns give the scoring rules defined in the text (the larger the better); "SE" denotes the squared prediction error of a point forecasts (the smaller the better). Superscript numbers are p-values (in percent) of the two-sided Giacomini-White test for equal predictive ability, using EW as a benchmark model.

Industrial Production

Model	**LS**	**QS**	**CRPS**	**SE**
		$h = 1$		
AR^s	$-1.04^{3.2}$	$0.51^{20.4}$	$-0.35^{0.9}$	$0.44^{1.1}$
AR^l	$-1.02^{2.1}$	$0.51^{31.8}$	$-0.33^{79.9}$	$0.39^{62.0}$
$\text{VAR}^s_{I,C}$	$-1.07^{0.6}$	$0.49^{1.0}$	$-0.36^{0.1}$	$0.46^{1.1}$
$\text{VAR}^l_{I,C}$	$-1.05^{6.7}$	$0.50^{3.1}$	$-0.34^{11.5}$	$0.41^{43.3}$
$\text{VAR}^s_{I,T}$	$-1.01^{22.2}$	$0.51^{51.8}$	$-0.34^{34.4}$	$0.41^{28.2}$
$\text{VAR}^l_{I,T}$	$-1.05^{5.8}$	$0.50^{12.0}$	$-0.34^{28.6}$	$0.41^{43.4}$
$\text{VAR}^s_{I,U}$	$-1.02^{9.7}$	$0.49^{1.3}$	$-0.35^{1.4}$	$0.43^{5.7}$
$\text{VAR}^l_{I,U}$	$-1.04^{0.9}$	$0.50^{1.4}$	$-0.34^{5.2}$	$0.41^{46.0}$
EW	-0.98	0.52	-0.33	0.40
(Rank)	(1)	(1)	(1)	(2)
		$h = 3$		
AR^s	$-1.05^{11.9}$	$0.49^{27.7}$	$-0.35^{2.9}$	$0.43^{2.4}$
AR^l	$-1.04^{9.6}$	$0.48^{15.8}$	$-0.33^{71.7}$	$0.37^{23.9}$
$\text{VAR}^s_{I,C}$	$-1.04^{21.1}$	$0.49^{18.4}$	$-0.35^{0.4}$	$0.42^{0.2}$
$\text{VAR}^l_{I,C}$	$-1.07^{7.9}$	$0.48^{3.1}$	$-0.34^{23.6}$	$0.41^{57.7}$
$\text{VAR}^s_{I,T}$	$-1.03^{31.9}$	$0.49^{37.7}$	$-0.34^{4.4}$	$0.42^{2.5}$
$\text{VAR}^l_{I,T}$	$-1.08^{6.1}$	$0.48^{10.2}$	$-0.34^{40.6}$	$0.40^{77.3}$
$\text{VAR}^s_{I,U}$	$-1.04^{15.3}$	$0.48^{13.7}$	$-0.35^{0.2}$	$0.43^{0.0}$
$\text{VAR}^l_{I,U}$	$-1.08^{3.2}$	$0.48^{7.7}$	$-0.34^{18.1}$	$0.40^{47.3}$
EW	-1.01	0.50	-0.34	0.40
(Rank)	(1)	(1)	(2)	(2)
		$h = 6$		
AR^s	$-1.14^{9.4}$	$0.47^{11.0}$	$-0.37^{4.7}$	$0.49^{8.8}$
AR^l	$-1.12^{5.3}$	$0.47^{9.4}$	$-0.35^{52.8}$	$0.42^{66.0}$
$\text{VAR}^s_{I,C}$	$-1.10^{56.1}$	$0.48^{46.9}$	$-0.35^{12.6}$	$0.43^{15.9}$
$\text{VAR}^l_{I,C}$	$-1.11^{9.4}$	$0.47^{34.7}$	$-0.35^{69.4}$	$0.41^{26.3}$
$\text{VAR}^s_{I,T}$	$-1.09^{80.2}$	$0.48^{56.8}$	$-0.35^{33.8}$	$0.43^{58.7}$
$\text{VAR}^l_{I,T}$	$-1.12^{7.0}$	$0.47^{19.7}$	$-0.35^{94.2}$	$0.41^{28.5}$
$\text{VAR}^s_{I,U}$	$-1.10^{48.0}$	$0.48^{39.6}$	$-0.35^{8.5}$	$0.44^{10.8}$
$\text{VAR}^l_{I,U}$	$-1.14^{1.1}$	$0.46^{5.3}$	$-0.35^{12.5}$	$0.43^{41.8}$
EW	-1.08	0.48	-0.35	0.43
(Rank)	(1)	(1)	(3)	(4)

Table 7: Same as Table 6, but for industrial production.

Treasury Bill Rate

Model	**LS**	**QS**	**CRPS**	**SE**
		$h = 1$		
AR^s	$-0.02^{58.9}$	$1.61^{0.4}$	$-0.12^{34.8}$	$0.04^{24.2}$
AR^l	$-0.23^{0.0}$	$1.11^{0.0}$	$-0.15^{0.0}$	$0.05^{0.0}$
$\text{VAR}^s_{T,C}$	$-0.02^{56.0}$	$1.61^{0.5}$	$-0.12^{58.4}$	$0.04^{5.6}$
$\text{VAR}^l_{T,C}$	$-0.26^{0.0}$	$1.07^{0.0}$	$-0.15^{0.0}$	$0.04^{1.6}$
$\text{VAR}^s_{T,I}$	$0.04^{99.4}$	$1.64^{0.1}$	$-0.12^{22.6}$	$0.04^{15.4}$
$\text{VAR}^l_{T,I}$	$-0.24^{0.0}$	$1.08^{0.0}$	$-0.14^{0.0}$	$0.04^{20.0}$
$\text{VAR}^s_{T,U}$	$-0.07^{45.0}$	$1.60^{0.7}$	$-0.12^{88.1}$	$0.04^{2.1}$
$\text{VAR}^l_{T,U}$	$-0.24^{0.0}$	$1.09^{0.0}$	$-0.14^{0.0}$	$0.04^{0.1}$
EW	0.04	1.49	-0.12	0.04
(Rank)	(1)	(5)	(5)	(1)
		$h = 3$		
AR^s	$-0.21^{45.7}$	$1.40^{5.3}$	$-0.13^{50.3}$	$0.05^{39.8}$
AR^l	$-0.32^{0.0}$	$0.99^{0.0}$	$-0.16^{0.0}$	$0.06^{0.1}$
$\text{VAR}^s_{T,C}$	$-0.21^{47.6}$	$1.40^{2.8}$	$-0.13^{52.1}$	$0.05^{39.1}$
$\text{VAR}^l_{T,C}$	$-0.35^{0.0}$	$0.96^{0.0}$	$-0.16^{0.0}$	$0.05^{21.8}$
$\text{VAR}^s_{T,I}$	$-0.11^{79.5}$	$1.44^{0.1}$	$-0.13^{5.9}$	$0.05^{77.1}$
$\text{VAR}^l_{T,I}$	$-0.34^{0.0}$	$0.98^{0.0}$	$-0.16^{0.0}$	$0.05^{47.4}$
$\text{VAR}^s_{T,U}$	$-0.21^{48.2}$	$1.42^{0.5}$	$-0.13^{51.8}$	$0.05^{32.1}$
$\text{VAR}^l_{T,U}$	$-0.34^{0.0}$	$0.98^{0.0}$	$-0.16^{0.0}$	$0.05^{77.7}$
EW	-0.08	1.30	-0.13	0.05
(Rank)	(1)	(5)	(5)	(3)
		$h = 6$		
AR^s	$-0.10^{99.2}$	$1.39^{0.7}$	$-0.13^{35.6}$	$0.05^{21.4}$
AR^l	$-0.34^{0.0}$	$0.97^{0.0}$	$-0.16^{0.0}$	$0.06^{0.5}$
$\text{VAR}^s_{T,C}$	$-0.09^{81.5}$	$1.39^{0.5}$	$-0.13^{23.7}$	$0.05^{30.7}$
$\text{VAR}^l_{T,C}$	$-0.36^{0.0}$	$0.96^{0.0}$	$-0.16^{0.0}$	$0.05^{72.0}$
$\text{VAR}^s_{T,I}$	$-0.09^{82.0}$	$1.39^{0.7}$	$-0.13^{27.7}$	$0.05^{44.4}$
$\text{VAR}^l_{T,I}$	$-0.36^{0.0}$	$0.96^{0.0}$	$-0.16^{0.0}$	$0.05^{80.6}$
$\text{VAR}^s_{T,U}$	$-0.09^{81.2}$	$1.38^{0.9}$	$-0.13^{24.4}$	$0.05^{37.2}$
$\text{VAR}^l_{T,U}$	$-0.36^{0.0}$	$0.96^{0.0}$	$-0.16^{0.0}$	$0.05^{95.8}$
EW	-0.10	1.28	-0.14	0.05
(Rank)	(4)	(5)	(5)	(2)

Table 8: Same as Table 6, but for the treasury bill rate.

Unemployment Rate

Model	**LS**	**QS**	**CRPS**	**SE**
		$h = 1$		
AR^s	$0.38^{2.1}$	$1.78^{0.4}$	$-0.09^{1.2}$	$0.03^{4.0}$
AR^l	$0.41^{14.9}$	$1.87^{61.5}$	$-0.09^{99.0}$	$0.02^{78.2}$
$\text{VAR}^s_{U,C}$	$0.37^{1.2}$	$1.79^{0.9}$	$-0.09^{0.2}$	$0.03^{0.2}$
$\text{VAR}^l_{U,C}$	$0.36^{3.3}$	$1.80^{0.7}$	$-0.09^{2.7}$	$0.03^{6.7}$
$\text{VAR}^s_{U,I}$	$0.40^{7.8}$	$1.80^{1.8}$	$-0.09^{2.8}$	$0.03^{7.9}$
$\text{VAR}^l_{U,I}$	$0.41^{11.4}$	$1.85^{30.2}$	$-0.09^{39.9}$	$0.02^{51.7}$
$\text{VAR}^s_{U,T}$	$0.39^{21.0}$	$1.86^{39.0}$	$-0.09^{11.0}$	$0.03^{7.1}$
$\text{VAR}^l_{U,T}$	$0.34^{3.8}$	$1.80^{3.7}$	$-0.09^{6.6}$	$0.03^{9.6}$
EW	0.44	1.89	-0.09	0.02
(Rank)	(1)	(1)	(1)	(2)
		$h = 3$		
AR^s	$0.38^{21.9}$	$1.80^{1.9}$	$-0.09^{6.9}$	$0.03^{14.7}$
AR^l	$0.40^{48.2}$	$1.86^{81.3}$	$-0.09^{60.7}$	$0.02^{42.1}$
$\text{VAR}^s_{U,C}$	$0.38^{8.4}$	$1.82^{3.7}$	$-0.09^{0.0}$	$0.03^{0.0}$
$\text{VAR}^l_{U,C}$	$0.37^{9.2}$	$1.84^{24.0}$	$-0.09^{37.9}$	$0.02^{53.6}$
$\text{VAR}^s_{U,I}$	$0.38^{8.0}$	$1.81^{2.0}$	$-0.09^{0.1}$	$0.03^{0.1}$
$\text{VAR}^l_{U,I}$	$0.38^{2.4}$	$1.84^{14.3}$	$-0.09^{79.3}$	$0.02^{42.4}$
$\text{VAR}^s_{U,T}$	$0.40^{30.5}$	$1.84^{24.5}$	$-0.09^{3.0}$	$0.03^{1.9}$
$\text{VAR}^l_{U,T}$	$0.36^{6.3}$	$1.83^{20.4}$	$-0.09^{36.3}$	$0.02^{60.3}$
EW	0.41	1.87	-0.09	0.02
(Rank)	(1)	(1)	(2)	(3)
		$h = 6$		
AR^s	$0.33^{9.5}$	$1.77^{2.5}$	$-0.09^{1.3}$	$0.03^{2.2}$
AR^l	$0.35^{29.3}$	$1.82^{68.5}$	$-0.09^{74.6}$	$0.03^{38.9}$
$\text{VAR}^s_{U,C}$	$0.35^{20.4}$	$1.80^{7.8}$	$-0.09^{0.1}$	$0.03^{0.0}$
$\text{VAR}^l_{U,C}$	$0.35^{15.2}$	$1.82^{60.8}$	$-0.09^{65.0}$	$0.03^{23.3}$
$\text{VAR}^s_{U,I}$	$0.34^{14.0}$	$1.79^{4.3}$	$-0.09^{0.0}$	$0.03^{0.0}$
$\text{VAR}^l_{U,I}$	$0.35^{7.4}$	$1.81^{22.5}$	$-0.09^{69.7}$	$0.03^{39.9}$
$\text{VAR}^s_{U,T}$	$0.36^{48.5}$	$1.80^{9.5}$	$-0.09^{1.8}$	$0.03^{2.6}$
$\text{VAR}^l_{U,T}$	$0.35^{6.5}$	$1.80^{16.6}$	$-0.09^{59.9}$	$0.03^{51.5}$
EW	0.37	1.83	-0.09	0.03
(Rank)	(1)	(1)	(3)	(5)

Table 9: Same as Table 6, but for the unemployment rate.