Combining Inflation Density Forecasts: Some Cross-Country Evidence

Christian Kascha and Francesco Ravazzolo*†

Norges Bank

Version: February 2008

Abstract

In this paper we empirically evaluate competing approaches for combining density forecasts. We compare combinations of density forecasts for CPI inflation using a suit of linear forecasting devices and various VARs with moving estimation windows to account for structural change. Three different data sets for the US, the UK and Norway are used. We find that several combination schemes improve over selecting the best model throughout the three data sets. Thus, it is safe to combine predictive densities. Furthermore, we find that some combination schemes that work well in point forecasting cannot be recommended for density combination.

JEL classification: C53, E37
Keywords: Forecast Combination, Density Forecasts, Inflation Forecasting

1 Introduction

The value of a point forecast can be increased by supplementing it with some measure of uncertainty. Interval and density forecasts are considered an important part of the communication from policymakers to the public. For example, the Bank of England as well as the central bank of Norway, Norges Bank, publish so-called fan-charts for inflation that supposedly communicate the banks' views on possible paths of future inflation. However, policymakers usually have a whole suit of forecast models at hand. While it is now established that the combination of individual forecasts may help to form a better consensus forecast when it comes to point forecasting, a similar conclusion has not been reached in the literature on density forecasting. In this paper, we therefore empirically evaluate competing approaches

*The views expressed in this paper are our own and do not necessarily reflect the views of Norges Bank.

†Norges Bank, Bankplassen 2, 0107 Oslo, Norway; Phone: +47 22 31 67 19, Fax: +47 22 42 40 62, E-mail: christian.kascha@norges-bank.no, francesco.ravazzolo@norges-bank.no.

for combining density forecasts. We compare combinations of density forecasts for inflation using a suit of linear forecasting devices and various VARs and compare the results over data sets for the US, the UK and Norway.

While the literature on combining point forecasts has reached a relatively mature state dating back to papers such as Bates & Granger (1969), the same cannot be said about our knowledge on density forecasting and particularly combination schemes of individual density forecasts. Timmermann (2006) provides an extensive summary of the literature on combining point forecasts.

The literature on density and interval forecasting is summarized in surveys of Tay & Wallis (2000) and Corradi & Swanson (2006a). See also Clemen, Murphy & Winkler (1995). Corradi & Swanson (2006a) is maybe the most comprehensive survey to date on the evaluation of single and multiple density forecasts.

Clements (2006) and Granger, White & Kamstra (1989) have considered combination of event and quantile forecasts. Density forecasting considering the whole distribution of the variable to be forecasted has only recently emerged in economics (see, e.g., Wallis 2005). In contrast, already Genest & Zidek (1986) provided a survey on density combination in meteorology.

Some combination schemes for density forecasts have been proposed that probably originate in their success in combining point forecasts. For example, equal weights have been suggested by Hendry & Clements (2004) and Wallis (2005). Granger & Jeon (2004) propose in a more general framework "thick modeling".

Bayesian approaches naturally lend themselves to density combination schemes. So it is maybe not surprising that various approaches have emerged in this field. For example, Min & Zellner (1993) propose simple combinations based on posterior odds ratios. Palm & Zellner (1992) propose a combination method that captures the full correlation structure between the forecast errors resulting from different models by explicitly modeling their dynamic interaction. Following Morris (1974), Morris (1977) and Winkler (1981), Hall & Mitchell (2004) consider an approach where competing density forecasts are combined by a "decision maker" who views these forecasts as data that is used to update a prior distribution. Bayesian model averaging (BMA) methods have been proposed by Leamer (1978) and Raftery, Madigan & Hoeting (1997). An empirical study on BMA is Jackson & Karlsson (2004). Geweke & Whiteman (2006) introduce the idea of predictive likelihood, and Eklund & Karlsson (2007) and Andersson & Karlsson (2007) use this method for forecast combination with Bayesian AR and VAR models. Garratt, Lee, Pesaran & Shin (2003) apply a Bayesian approach to the UK economy. Also Mitchell & Hall (2005) suggest weights derived in a Bayesian framework.

From a classical perspective, weights based on the Kullback-Leibler Information Criterion have been proposed by both Amisano & Giacomini (2007) and Hall & Mitchell (2007). This criterion offers a way of measuring predictive accuracy and points to combination schemes

seeking to minimize the distance between estimated predictive densities and the true but unknown density of the variable to be predicted. This is achieved by considering logarithmic scoring rules that reward models which on average give higher probability to events which have actually occurred.

While Jore, Mitchell & Vahey (2007) provide some evidence on the performance of the weighting scheme proposed by Hall & Mitchell (2007) relative to equal weights and the pairwise equal averaging method of Clark & McCracken (2007), our knowledge on when and why predictive density combinations work is still limited. As Hall & Mitchell (2007) state: "It is important to try to build up both an increased understanding and an empirical consensus about the circumstances in which density forecast combination works." Taking inflation density forecasting as a relevant example, we ask how predictive density combinations perform relative to individual density forecasts and selecting the best performing model ex-ante. How large are the gains from combining? Are some density combination schemes dominated by others? How do our results vary over the three samples? Finally, we compare the results on density combination with the results on point-forecast combination measured in terms of root mean squared prediction error.

Our results show that recursive log-score weights, trimming the $k\%$ worst models and the pairwise equal weight scheme give superior results than selecting ex-ante the best individual model in all the three samples. Furthermore, the obtained performance from combining is close to the result for the ex-post best individual model for the US application. The trimming approach provides the highest log-score in the Norwegian exercise. Equal weights, which provide accurate point forecasts, give, on contrary, always lower log-scores than selecting the best individual model ex-ante. This finding suggests that the equal weight scheme is inadequate for density forecasts. Bayesian model averaging confirms to have trimming effect, but its success depends on the measure used to derive model probabilities and how appropriate this measure is for the application of interest.

The rest of the paper is organized as follows. In section two we discuss the evaluation of density forecasts and combining predictive densities. In section three we describe the data and the suit of density forecast models. Section four contains the results of the out-of-sample experiment. Section five concludes.

2 Evaluating and Combining Predictive Densities

2.1 Evaluating Predictive Densities

The question of how to measure the accuracy of density forecasts has recently received a lot of attention in the theoretical literature. Corradi & Swanson (2006a) provide an extensive survey. This question is decisive because it is central to how we design density combination schemes (Hall & Mitchell 2007). In this paper we compare multiple models and combination

schemes that are misspecified and sometimes nested. This fact complicates matters by some degree.

One branch of the literature is concerned with testing whether predictive densities are correctly specified (Bierens 1982, Bierens & Ploberger 1997). These tests require the assumption of correct specification of the density forecast under the null using all the relevant information (e.g., Diebold, Gunther & Tay 1998, Bai 2003) or conditional on a given information set (Corradi & Swanson 2003). Unfortunately, the models in our suit do not satisfy any of these assumptions.

Another branch is concerned with model evaluation of multiple, possibly misspecified models. One possibility is to evaluate density forecasts in terms of their implied economic value (Granger & Pesaran 2000, Clements 2004). This strategy makes a lot of sense in areas such as financial econometrics but is less meaningful for policymakers such as central banks. Therefore we stick to statistical measures. Two approaches have been considered in the recent literature. One is based on a distributional analog of mean squared error (Corradi & Swanson 2004, 2006b), the other is based on the Kullback-Leibler Information Criterion (KLIC) (Kitamura 2002, Amisano & Giacomini 2007, Hall & Mitchell 2007).

The measure of distributional accuracy introduced by Corradi & Swanson (2004, 2006b) is attractive because of its analogy to the usual mean squared error norm in point forecasting. Given a benchmark density function, a norm over a set of possible density forecasts is defined in a straightforward manner taking the expectation of the squared, point-wise difference between a candidate density and the benchmark density over all possible outcomes of the variable to be forecasted. One problem is the dependence on a benchmark which might be difficult to justify in our case unless one uses a nonparametric estimate as in Li & Tkacz (2006). Given the short sample sizes at hand, this option does not appear convenient in the present context.

On the other hand, measures based on the well-known KLIC can circumvent this problem. The KLIC is a sensible measure of accuracy, since it chooses the model which on average gives higher probability to events that have actually occurred. Specifically, the KLIC distance between the true density f_t and some candidate density f_t^c of a random variable Y_t is defined as

$$\begin{aligned} \text{KLIC}_t &= \int f_t(y_t) \ln \frac{f_t(y_t)}{f_t^c(y_t)} dy_t \\ &= E[\ln f_t(y_t) - \ln f_t^c(y_t)]. \end{aligned}$$

Under some regularity conditions, a consistent estimate can obtained from the average of the sample information, $y_1, \ldots, y_T$, on f_t and f_t^c:

$$\overline{KLIC} = \frac{1}{T} \sum_{t=1}^{T} [\ln f_t(y_t) - \ln f_t^c(y_t)].$$

Even though we do not know the true density, we can still compare multiple densities, $f^c_{t,i}$ for $i = 1, \ldots, N$. It is sufficient to consider only the latter term in the above sum,

$$-\frac{1}{T}\sum_{t=1}^{T} \ln f^c_{t,i}(y_t), \tag{1}$$

for all i and to choose the model for which the expression in (1) is minimal.

Turning to density forecasts, let $f^c_{t+h,t}$ denote a prediction of the density for Y_{t+h}, possibly obtained by combination of individual density forecasts and conditional on information up to date t. Let y_{t+h} be the realization of Y_{t+h} and suppose that T^e h-step-ahead-forecasts have been obtained, starting at time T^s and given a total number of T observations. A measure of out-of-sample performance is the average logarithmic score or "log-score"

$$\ln S := \frac{1}{T^e}\sum_{t=T^s}^{T-h} \ln f^c_{t+h,t}(y_{t+h}). \tag{2}$$

Models or combination schemes that are associated with a high average log-score are approximating well the unknown true density in terms of KLIC (Hall & Mitchell 2007).

For the *iid* case, Vuong (1989) suggests a likelihood ratio test for choosing the conditional density model that is close to the true density in terms of KLIC. The tests was extended by Amisano & Giacomini (2007) to cover the case of dependent observations. Also Kitamura (2002) employs a KLIC-based approach to select between misspecified models. Measures in terms of the KLIC also have a Bayesian interpretation as the KLIC-best model is also the model with the highest posterior probability, as shown by Fernández-Villaverde & Rubio-Ramirez (2004). Hall & Mitchell (2007) propose to use the KLIC distance between the combined density forecast and the true but unknown density of the variable that is forecasted. Practically, computation of the KLIC distance does not require the estimation of any statistical model. Thus, it can measure forecasts obtained from informal sources as well and is our preferred measure of predictive accuracy.

2.2 Combining Predictive Densities

There are two elementary choices in combining densities. One is the way of aggregation and the other is the construction of the weights. Possible ways of aggregation are described in Genest & Zidek (1986). We focus here on the "linear opinion pool" of N competitive forecast densities of the same event:

$$f^c_{t+h,t}(y_{t+h}) \quad = \quad \sum_{i=1}^{N} \omega_{t+h,t,i} f_{t+h,t,i}(y_{t+h}), \tag{3}$$

where the collection of probability forecasts is $\{f_{t+h,t,1}, \ldots, f_{t+h,t,N}\}$. The weights have to be a convex linear combination, that is, $0 \leq \omega_{t+h,t,i} \leq 1$ and $\sum_{i=1}^{N} \omega_{t+h,t,i} = 1$ for all $i = 1, \ldots, N$ such that the resulting combination is indeed a density function. There are also other schemes such as the generalized opinion pool and a logarithmic combination of densities (Genest & Zidek 1986). Both are interesting alternatives, but we focus on the linear opinion pool for simplicity.

For the derivation of the N probability forecasts we apply Bayesian inference on individual models in section 3. This choice simplifies our computation as in Bayesian forecasting it is natural to compute predictive densities, which thereby we define as forecast densities, and we can use standard results to compute density forecasts from the set of individual models in 3. The conditional forecast density of a future random variable Y_{t+h} given the data up to time t, $\Omega_t = \{y_s, x_s | s \leq t\}$, and model i is defined as

$$f_{t+h,t,i}(y_{t+h}) = \int f_{t+h,t,i}(y_{t+h}|\theta_i) p(\theta_i|\Omega_t) d\theta_i, \tag{4}$$

where $f_{t+h,t,i}(y_{t+h}|\theta_i)$ is the conditional forecast density of Y_{T+h} given Ω_t, model parameters θ_i, and model i; $p(\theta_i|d_t)$ is the posterior density for parameter θ_i. For the construction of the weights, we consider several recent proposals in the emerging literature as well as the empirical evidence on the combination of point forecasts (Timmermann 2006).

Equal weights (EW): Equal weights are used in the aggregation of the forecasts in the Survey of Professional Forecasters to publish a combined density forecast for inflation. Equal weights for combining densities have also been proposed in the literature by Hendry & Clements (2004) and Wallis (2005). Formally, $\omega_{t+h,t,i} = 1/N$ for all t, h, i.

Recursive log-score weights (RLSW): If we are measuring density fit by the logarithmic score, then it is only natural to base the construction of combination weights on past out-of-sample forecast performance measured in the same way. A promising candidate combination scheme are recursive log-score weights as proposed in Jore et al. (2007). The weights for the h-step ahead density combination take the form

$$w_{t+h,t,i} = \frac{\exp[\sum_{\tau=\underline{t}}^{t-h} \ln f_{\tau+h,\tau,i}(y_{\tau+h})]}{\sum_{i=1}^{N} \exp[\sum_{\tau=\underline{t}}^{t-h} \ln f_{\tau+h,\tau,i}(y_{\tau+h})]}, \tag{5}$$

where $\underline{t}$ is the beginning of the evaluation period and is taken as fixed.

Trimming (TRIMW): Trimming the set of models by throwing away the k% worst models and assigning equal weights to the remaining models is a popular way of improving EW forecast combinations (Granger & Jeon 2004, Timmermann 2006). However, as Granger &

Jeon (2004) state "[...], this is more of a pragmatic folk-view than anything based on a clear theory.". A more practical concern is the choice of k and the evaluation criterion. The evaluation criterion is past out-of-sample performance in terms of the average log-score. At each point of time and for each horizon we choose the trimming parameter k computing the average log-score for a number of alternatives

$$\ln S_{t,k,h} = \frac{1}{t-h-\underline{t}+1} \sum_{\tau=\underline{t}}^{t-h} \ln f^{c,k}_{\tau+h,\tau,i}(y_{t+h}) \tag{6}$$

where $\underline{t}$ is the beginning of the evaluation period and $f^{c,k}$ is the density combination obtained by trimming the $k\%$ worst models. Then, the k is chosen that yielded the highest out-of-sample average log-score in the past. The set of possible ks is $\{8, 16, \ldots, 62\}$, which corresponds to throwing away $1, 2, \ldots, 8$ density forecasts, respectively.

Pairwise equal weights (PEW): Pairwise equal weights (PEW) is actually a special case of trimming and has been proposed by Clark & McCracken (2007) for point forecasts. They propose to combine the forecasts of one univariate model and one multivariate model with equal weights. We consider a slight modification of this strategy by choosing at each point of time two models based on past performance of the models' density forecasts in terms of the average log-score.

Bayesian Model Averaging

Bayesian approaches have been widely used to construct forecast combinations, see, e.g., Leamer (1978), Hodges (1987), Draper (1995), Min & Zellner (1993) and Strachan & van Dijk (2007). In this approach one does not estimate regression weights and uses those to compute (density) forecasts, but the combination weights are based on the posterior probability for any individual model. The predictive density accounts then for model uncertainty by averaging over the probabilities of individual models. We propose two BMA schemes: the original one proposed in an empirical application by Madigan & Raftery (1994), and a more recent one discussed in Geweke & Whiteman (2006).

BMA using marginal likelihood (BMAW): The forecast density of Y_{t+h} given Ω_t is computed by averaging over the conditional forecast densities given the individual models with the posterior probabilities of these models as weights:

$$f_{t+h,t}(y_{t+h}) = \sum_{i=1}^{N} f_{t+h,t,i}(y_{t+h}) P(i|\Omega_t), \tag{7}$$

where $f_{t+h,t,i}(y_{t+h})$ is the conditional predictive density given Ω_t and model i; $P(i|\Omega_t)$ is the posterior probability for model i, defined as

$$P(i|\Omega_t) = \frac{p(y|i)p(i)}{\sum_{j=1}^{N} p(y|j)p(j)}, \tag{8}$$

where $y = \{y_s\}_{s=1}^{t}$; $p(i)$ is the prior density for model i and $p(y|i)$ is the marginal likelihood for model i given by

$$p(y|i) = \int p(\theta_i|\Omega_t, i)p(\theta_i)d\theta_i. \tag{9}$$

$p(\theta_i)$ is the prior density for the parameter θ_i of model i. The integral in equation (9) can be evaluated analytically in the case of linear models, but not for more complex forms. Chib (1995), e.g., has derived a method to compute the above expression also for some nonlinear examples. Proper priors for θ_i are usually applied, otherwise the Bartlett paradox may hold and models with less parameters would be strictly preferred.

BMA using predictive likelihood (BMAPLW): Geweke & Whiteman (2006) propose a BMA scheme based on the idea that a model is as good as its predictions. The predictive density of Y_{t+h} conditional on Ω_t has the same form as equation (7), but the posterior density of model i conditional on Ω_t is now computed as:

$$P(i|\Omega_t) = \frac{f_{t,t-h,i}(y_t)p(i)}{\sum_{j=1}^{n} f_{t,t-h,j}(y_t)p(j)}, \tag{10}$$

where $f_{t,t-h,i}(y_t)$ is as in (5). In Bayesian terminology, this density is defined as the predictive likelihood for model i. As in de Pooter, Ravazzolo & van Dijk (2007) and Ravazzolo, van Dijk & Verbeek (2007) we compute the predictive density for quarter t using information until quarter $t-h$ and we evaluate the *realized* value for time t using the same density. The resulting probability is then applied to compute the weight for model i in constructing the forecast for $t+h$ made at time t. If we evaluated the predictive likelihood over an expanding window of forecasts, this approach would coincide with the recursive log-score weights.

3 The Data and the Model Suite

We take inflation density forecasting as a relevant example to evaluate different ways of combining predictive densities. In order to obtain an intuition for the sample dependence of our and other results we compare ways of combining density forecasts over three different data sets. Let p_t denote some price level index in quarter t. We are interested in forecasting quarter-to-quarter inflation measured by the quarterly log change, $\Delta_1 p_t = \ln p_t - \ln p_{t-1}$. We consider CPI indices for the US and the UK and the Norwegian core CPI index. The set of potential predictors contains a quarterly money measure, a three month Treasury Bill yield

and a quarterly output measure. We focus on core CPI for Norway as energy prices have a dominant role on the Norwegian CPI index. Norwegian energy prices in turn are affected largely by weather conditions. We use M2 as a money measure in all three countries, and we use real output as a measure of US GDP. Quarterly real GDP series are available for the other two countries.

We collect US CPI, GDP and M2 data from the Federal Reserve Bank of Philadelphia's Real time Data Set for Macroeconomists, US interest rates from the Fred database, UK CPI, interest rates and money from the OECD database, and UK GDP from EUROSTAT. Norwegian data is collected from Norges Banks database.[1] We use seasonally unadjusted series and even though we have real time data for US and Norwegian GDP we simply abstract from the real time aspect of the data and use the latest available vintage.

Data is available over different sample periods for the three countries and we select the longest sample period we have for each country. We consider sample periods that run from 1960 Q1 to 2007 Q3 for US data, from 1978 Q1 to 2007 Q2 for UK data, and from 1979 Q2 to 2007 Q3 for Norwegian data. The evaluation periods start with the forecasts undertaken in 1976 Q2, 1994Q2 and 1995 Q4. Thus, there are 126, 53 and 48 evaluated forecasts for the US, the UK and Norway, respectively.

Univariate Models: The univariate models may be justified as simple "forecasting devices" as in Clements & Hendry (2006). These simple models are included to insure against all sorts of structural breaks. They also present different assumptions about the orders of integration of the price level series. These models can be quite serious forecasting devices as pointed out by, e.g., Castle & Hendry (2007).

Random Walk in $\Delta_4 \ln p_t$ (RWD4):

The model is given by

$$\Delta_4 \ln p_t = \Delta_4 \ln p_{t-1} + \varepsilon_t,$$

where $\Delta_4 \ln p_t := \ln p_t - \ln p_{t-4}$ and the forecast at horizon h is $\Delta_4 \widehat{\ln p}_{T+h|T} = \Delta_4 \ln p_T$.

Random Walk in $\Delta_1 \ln p_t$ (RWD1):

The model is given by

$$\Delta_1 \ln p_t = \Delta_1 \ln p_{t-1} + \varepsilon_t,$$

and forecast can be obtained directly.

[1] All data are available upon request to the authors.

Random Walk in $\Delta_1\Delta_4 \ln p_t$ (RWD1D4):

The model is given by

$$\Delta_1\Delta_4 \ln p_t = \Delta_1\Delta_4 \ln p_{t-1} + \varepsilon_t,$$

where $\Delta_1\Delta_4 \ln p_t := \Delta_1(\ln p_t - \ln p_{t-4}) = (\ln p_t - \ln p_{t-4}) - (\ln p_{t-1} - \ln p_{t-5})$.

AR1D4:

The model is a simple AR(1) in fourth differences

$$\Delta_4 \ln p_t = \mu + \alpha_1\Delta_4 \ln p_{t-1} + \varepsilon_t,$$

where, for estimation purposes, only the last 20 observations are used.

AR1D1:

The model is a simple AR(1) in first differences

$$\Delta_1 \ln p_t = \mu + \alpha_1\Delta_1 \ln p_{t-1} + \varepsilon_t,$$

where, for estimation purposes, only the last 20 observations are used.

Vector Autoregressive Models: The model suite also contains vector autoregressive models of the form

$$\mathbf{y}_t = \mu + A_1\mathbf{y}_{t-1} + \ldots + A_p\mathbf{y}_{t-p} + u_t,$$

where $\mathbf{y}_t$ is a $(K \times 1)$ random vector, μ, $A_1, \ldots,$ A_p are constant coefficient matrices of suitable dimension, u_t is the error term and the lag length is denoted by p.

We consider different VARs that contain variables usually considered in the literature on forecasting inflation. All VARs are estimated with 2 lags using the last 50 observations to allow for structural change. The VARs are unrestricted. Therefore, the models can be distinguished by the components in $\mathbf{y}_t$ alone. We use eight different VARs whose components are given in table 1. In the table, $M2_t$ denotes the M2 money measure, i_t a short-term interest rate and y_t is a quarterly output measure.

In estimation we use conjugate diffuse priors. That is, we basically use uninformative priors and the resulting estimates are therefore not very different from their classical counterparts. The reason we apply Bayesian methods at the estimation stage is that we use Bayesian combination methods later on and we would like to apply different combination schemes to the same predictive densities. We refer to Koop (2003) for estimation details on univariate models and de Pooter et al. (2007) for derivations on VAR models. In general, if y_t denotes the variable to be forecasted by model i, the predictive density for the one-step-ahead forecast

Table 1: Description of employed VARs

VAR-Nr.	Variables
1	$(\Delta_1 \ln p_t,\ \Delta_1 \ln M2_t)'$
2	$(\Delta_1 \ln p_t,\ \Delta_1 i_t)'$
3	$(\Delta_1 \ln p_t,\ \Delta_1 i_t,\ \Delta_1 \ln y_t)'$
4	$(\Delta_1 \ln p_t,\ \Delta_1 \ln M2_t,\ \Delta_1 \ln y_t)'$
5	$(\Delta_4 \ln p_t,\ \Delta_4 \ln M2_t)'$
6	$(\Delta_4 \ln p_t,\ \Delta_4 i_t)'$
7	$(\Delta_4 \ln p_t,\ \Delta_4 i_t,\ \Delta_4 \ln y_t)'$
8	$(\Delta_4 \ln p_t,\ \Delta_4 \ln M2_t,\ \Delta_4 \ln y_t)'$

follows a $t(\mu_i,\ \Sigma_i,\ v)$-distribution given by

$$\begin{aligned} f_{t+1,t,i}(y_{t+1}, \mu_i, \Sigma_i, \nu_i) &= c_i\ |\Sigma_i|^{-1/2}[\nu_i + (y_{t+1} - \mu_i)\Sigma_i^{-1}(y_{t+1} - \mu_i)]^{-\frac{\nu_i + k_i}{2}} \\ c_i &= \frac{v_i^{v_i/2}\Gamma(\frac{\nu_i + k_i}{2})}{\pi^{k_i/2}\Gamma(\frac{v_i}{2})} \end{aligned}$$

for a $(k_i \times 1)$ vector μ_i, a positive definite matrix Σ_i $(k_i \times k_i)$ and a positive scalar v_i, where k_i denotes the number of variables in the model. In case of the VARs $k_i > 1$ and the marginal density for an element $y_{t+1,j}$ of y_{t+1} is $t(\mu_{i,j},\ \Sigma_{i,jj},\ v)$, where $\mu_{i,j},\ \Sigma_{i,jj}$ denote the jth entry in μ_i and the (j,j) element in Σ_i, respectively (Koop 2003).

Selection (SELEC): Comparing the performance of the combination methods to the performance of each individual model is interesting. However, this kind of comparison is less informative about the actual forecasting performance that could be obtained in real-time because it is essentially ex-post. In practice, a forecaster has to choose ex-ante which forecasting model to employ. We therefore compare the combination methods to ex-ante selection.[2] This can be done in several ways. It is however natural to assume that a forecaster, *if* he has to select one model, chooses the model that performed best in the past. Since we are interested in predictive densities, the relevant criterion is here the past performance of the models in terms of the average log-score. Note that (i) the way we select models should be closely related to the standard AIC criterion based on the predictive likelihood and that (ii) the above described PEW scheme is close to selection.

[2]Another interesting issue that is not explored in this paper is the following: The model suit is taken as given but of course the suit is chosen because we have an idea which models might work in the present context. This idea is mainly based on evaluating the models ex-post - using the available data. Thus, we should view the set of models as random. A more sophisticated comparison would therefore model the evolution of the model set as well. This is however left for future research.

4 Out-of-Sample Experiment

The results of the out-of-sample evaluation are summarized in figures 1 - 3 and in table 2. We focus here on one-step-ahead density forecasting for simplicity. Generating forecast densities for longer forecast horizons requires simulation of the predictive densities and would therefore increase the computational burden quite quickly. We are currently working on an extension to multi-step-ahead forecasts. Figures 1 - 3 show the out-of-sample forecasting performance of the individual models, the combination schemes and model selection for the three data sets. Out-of-sample forecasting performance is measured both in terms of the average log-score and RMSPE. In the figures we display the negative average log-score such that models which are close to the $(0, 0)$ coordinate perform well in terms of both measures. Table 2 tabulates exactly the same information that has been used to generate the figures.

The results for the individual models show that while there is a close relation between a model's average log-score and its RMSPE for most of the models, this relation might break down in important cases. In the case of the UK and Norway, the AR1D4 model is performing miserably in terms of the average log-score but is quite competitive in terms of RMSPE. As expected, the forecasting performance of the individual models varies considerably over data sets. While the RWD1D4 model is performing relatively well in terms of log-score and RMSPE for the US, the same cannot be said for the other two data sets. This might be explained by the fact that the US sample is much longer than the other two and contains periods of high volatility such as the 1970s. The VARs can generate surprisingly good predictive densities, in particular bivariate VARs containing log prices and interest rates in first (VAR2) or fourth (VAR6) differences do well for all three data sets. Considering higher-dimensional VARs does not seem to pay off in our setting. In the case of the US, the VAR2 is the best model in terms of average log-score. The best univariate model (RWD4) in terms of average log-score dominates however all the VARs in the case of the UK and Norway.

As in Hall & Mitchell (2007), the results are partially disappointing for an exercise where the forecaster trying to select in real-time the best model among a set of competitive specifications. This approach, SELECT, provides occasionally poor statistics, with a reduction of average log-score of roughly 15% and 50% compared to the best individual model for the US and the UK, respectively. Moreover, in both cases there are at least two individual models that perform better than the SELECT approach. Results are qualitatively similar in terms of point forecast accuracy. The evidence is different for Norway, where the log-score of the SELECT approach is equal to the best model. The explanation is that the RWD4 specification provides the best forecasts at each point of time for Norway, meaning that it is always selected.

Focusing on the lower part of table 2, findings for combination schemes are very intriguing. Combining forecasts is the best forecasting strategy in several cases, and it is always a "safe" approach to minimize forecast errors. There are however important difference among

evaluation criteria and data sets. The RLSW, TRIMW and PEW schemes provide higher log-scores and lower RMSPEs than the SELECT approach for the US and Norway. TRIMW has the higher log-score among the three approaches, and in Norway it has the highest log-scores among all models and combinations. TRIMW does also well in term of RMSPE. The log-score associated to TRIMW is however lower than the log-score of the SELECT strategy for UK data. RLSW and PEW schemes are more consistent over the three applications, with results for RLSW better in US data, but opposite evidence for UK and Norway. Therefore, opposite to Jore et al. (2007), which focus only on US data and have a smaller set of univariate models, our results revaluate PEW and TRIMW schemes. Note, however, that they use slightly different evaluation criteria. The explanation probably relates to our stronger attention to structural instability. As Castle & Hendry (2007) describe, our set of univariate devices is robust to various forms of instability, which is not always the case in Jore et al. (2007). Moreover, our selection of the two or $N - k$ best individual models to use in the PEW or TRIMW respectively, is done ex-ante at each point of time, allowing for different combinations over the sample period. Jore et al. (2007) use the same two individual models over the full forecasting sample, ignoring that it may be possible to predict that some models perform well over some periods but not in others as Hall & Mitchell (2007) also notice.

Bayesian model averaging perform poorly with US data, but gives quite accurate forecasts for UK and Norwegian density and point inflation forecasts. In particular, results are very promising for the scheme BMAW applied to the UK. As Hall & Mitchell (2007) also conclude, for this database using in-sample information to derive model weights is a better strategy that computing weights using past performance. BMAPLW has lower log-score statistics than those given by the SELECT approach for all the three countries. BMAPLW assigns weights by computing the realized predictive density of the last forecast. This seems less adequate than computing weights over an expanding set of realized probabilities as it is done for the RLSW or discarding the worse models using the same statistics.

We finally focus on the performance of the EW scheme. Clark & McCracken (2007) shows that equal weights give often more accurate point forecasts than more complex combinations schemes. We find similar evidence in term of RMSPE. However, as in Jore et al. (2007), results are very different in terms of average log-score. The average log-scores of the EW scheme is lower than those obtained by SELECT and other averaging schemes in all three exercises. Furthermore, differences are often quite substantial, indicating that this averaging scheme is not adequate for density forecasts.

5 Conclusion

This paper proposes to expand the research on combining inflation density forecasts by evaluating several averaging schemes over three different data sets. Equal weights, recursive log-

score weights, trimming, pairwise equal weights and two Bayesian model averaging schemes are proposed to combine density forecasts from a set of univariate devices and multivariate VARs for US, UK and Norwegian inflation. Results are evaluated in terms of average log-score. We find that combination schemes do not always beat the ex-post best individual models, but recursive log-score weight, trimming and pairwise equal weight schemes always outperform a strategy where the best individual model is selected ex-ante at each point of time. Bayesian model averaging based on marginal likelihood also provides promising results, and it seems more appropriate in exercises where the previous schemes perform poorly. Finally, the equal weight scheme gives always worse results than other combination methods and selecting ex-ante the best model in all three exercises, confirming that this scheme is not adequate for density forecasting as it is for point forecasting.

We think our findings can provide some directions to future applied research on density forecasts. For example, recursive weights and Bayesian model averaging methods can be combined with trimming. The worst k models can first be trimmed out, then estimated weights can be assigned to the remaining $N - k$ models. Other combination schemes where weights are estimated using different criteria can be exploited, in particular considering the poor performance of equal weights.

References

Amisano, G. & Giacomini, R. (2007), 'Comparing density forecasts via weighted likelihood ratio tests.', *Journal of Business & Economic Statistics* **25**, 177–190.

Andersson, M. & Karlsson, S. (2007), Bayesian forecast combination for VAR models. Unpublished manuscript, Sveriges Riksbank.

Bai, J. (2003), 'Testing parametric conditional distributions of dynamic models.', *Review of Economics and Statistics* **85**, 531–549.

Bates, J. M. & Granger, C. W. J. (1969), 'Combination of forecasts', *Operational Research Quarterly* **20**(4), 451–468.

Bierens, H. (1982), 'Consistent model-specification tests.', *Journal of Econometrics* **20**, 105–134.

Bierens, H. J. & Ploberger, W. (1997), 'Asymptotic theory of integrated conditional moments tests.', *Econometrica* **65**, 1129–1151.

Castle, J. L. & Hendry, D. F. (2007), Forecasting uk inflation: The roles of structural breaks and time disaggregation. University of Oxford, Department of Economics, Discussion Paper.

Chib, S. (1995), 'Marginal likelihood from the gibbs output', *Journal of American Statistical Association* **90**, 972–985.

Clark, T. E. & McCracken, M. W. (2007), Averaging forecasts from VARs with uncertain instabilities. Revision of Federal Reserve Bank of Kansas City Working Paper 06-12.

Clemen, R. T., Murphy, A. H. & Winkler, R. L. (1995), 'Screening probability forecasts: Contrasts between choosing and combining.', *International Journal of Forecasting* **11**, 133–145.

Clements, M. P. (2004), 'Evaluating the bank of england density forecasts of inflation.', *Economic Journal* **114**, 844–866.

Clements, M. P. (2006), 'Evaluating the survey of professional forecasters probability distributions of expected inflation based on derived event probability forecasts.', *Empirical Economics* **31**, 49–64.

Clements, M. P. & Hendry, D. F. (2006), *Handbook of Economic Forecasting*, Elsevier, chapter 9.

Corradi, V. & Swanson, N. (2003), 'Bootstrap conditional distribution tests in the presence of dynamic misspecification.', *Journal of Econometrics* **133**(2), 779–806.

Corradi, V. & Swanson, N. (2004), A test for comparing multiple misspecified conditional distributions. Working Paper, Rutgers University.

Corradi, V. & Swanson, N. (2006a), *Handbook of Economic Forecasting*, Elsevier, chapter 2.

Corradi, V. & Swanson, N. (2006b), 'Predictive density and conditional confidence interval accuracy tests.', *Journal of Econometrics* **135**, 187–228. 1-2.

de Pooter, M., Ravazzolo, F. & van Dijk, D. (2007), 'Predicting the term structure of interest rates', *Working paper, Tinbergen Institute* .

Diebold, F. X., Gunther, T. & Tay, A. S. (1998), 'Evaluating density forecasts with applications to finance and management.', *International Economic Review* **39**, 863–883.

Draper, D. (1995), 'Assessment and propagation of model uncertainty', *Journal of the Royal Statistical Society Series B* **56**, 45–98.

Eklund, J. & Karlsson, S. (2007), 'Forecast combination and model averaging using predictive measures', *Econometric Reviews* **26**, 329–362.

Fernández-Villaverde, J. & Rubio-Ramirez, J. F. (2004), 'Comparing dynamic equilibrium models to data', *Journal of Econometrics* **123**, 153–187.

Garratt, A., Lee, K., Pesaran, M. H. & Shin, Y. (2003), 'Forecast uncertainties in macroeconometric modelling: An application to the UK economy.', *Journal of the American Statistical Association* **98**, 829–838.

Genest, C. & Zidek, J. (1986), 'Combining probability distributions: A critique and an annotated bibliography.', *Statistical Science* **1**, 114–148.

Geweke, J. & Whiteman, C. (2006), Bayesian forecasting, *in* G. Elliot, C. Granger & A. Timmermann, eds, 'Handbook of Economic Forecasting', North-Holland.

Granger, C. W. J. & Jeon, Y. (2004), 'Thick modeling.', *Economic Modelling* **21**, 323–343.

Granger, C. W. J. & Pesaran, M. H. (2000), 'Economic and statistical measures of forecast accuracy.', *Journal of Forecasting* **19**, 537–560.

Granger, C. W. J., White, H. & Kamstra, M. (1989), 'Interval forecasting: An analysis based upon ARCH-quantile estimators.', *Journal of Econometrics* **40**, 87–96.

Hall, S. G. & Mitchell, J. (2004), Density forecast combination. National institute of economic and social research discussion paper, No. 249.

Hall, S. G. & Mitchell, J. (2007), 'Combining density forecasts.', *International Journal of Forecasting* **23**, 1–13.

Hendry, D. F. & Clements, M. P. (2004), 'Pooling of forecasts.', *Econometrics Journal* **7**, 1–31.

Hodges, J. (1987), 'Uncertainty, policy analysis and statistics', *Statistical Science* **2**, 259–291.

Jackson, T. & Karlsson, S. (2004), 'Finding good predictors for inflation: A bayesian model averaging approach.', *Journal of Forecasting* **23**, 479–498.

Jore, A. S., Mitchell, J. & Vahey, S. P. (2007), Combining forecast densities from VARs with uncertain instabilities. Norges Bank, NIESR and RBNZ, working paper.

Kitamura, Y. (2002), Econometric comparisons of conditional models. Working Paper, University of Pennsylvania.

Koop, G. (2003), *Bayesian Econometrics*, Wiley.

Leamer, E. (1978), *Specification Searches.*, Wiley, Oxford.

Li, F. & Tkacz, G. (2006), 'A consistent bootstrap test for conditional density functions with time-dependent data.', *Journal of Econometrics* **127**, 863–886.

Madigan, D. M. & Raftery, A. E. (1994), 'Model selection and accounting for model uncertainty in graphical models using occam's window', *Journal of the American Statistical Association* **89**, 1335–1346.

Min, C. K. & Zellner, A. (1993), 'Bayesian and non-bayesian methods for combining models and forecasts with applications to forecasting international growth rates.', *Journal of Econometrics* **56**, 89–118.

Mitchell, J. & Hall, S. G. (2005), 'Evaluating, comparing and combining density forecasts using the klic with an application to the bank of england and nieser "fan" charts of inflation.', *Oxford Bulletin of Economics and Statistics* **67**, 995–1033.

Morris, P. (1974), 'Decision analysis expert use.', *Managment Science* **20**, 1233–1241.

Morris, P. (1977), 'Combining expert judgments: A bayesian approach.', *Managment Science* **23**, 679–693.

Palm, F. C. & Zellner, A. (1992), 'To combine or not to combine?', *Journal of Forecasting* **11**, 687–701.

Raftery, A. E., Madigan, D. & Hoeting, J. A. (1997), 'Bayesian model averaging for linear regression models.', *Journal of the Amercian Statistical Association* **92**, 179–191.

Ravazzolo, F., van Dijk, H. K. & Verbeek, M. (2007), 'Predictive gains from forecast combination using time-varying model weight', *Econometric Institute Report 2007-26* .

Strachan, R. & van Dijk, H. K. (2007), 'Bayesian model averaging in vector autoregressive processes with an investigation of stability of the us great ratios and risk of a liquidity trap in the usa, uk and japan', *Econometric Institute Report 2007-09* p. 47.

Tay, A. S. & Wallis, K. F. (2000), 'Density forecasting: A survey.', *Journal of Forecasting* **19**, 235–254.

Timmermann, A. (2006), *Forecast Combinations*, Elsevier, chapter 4.

Vuong, Q. (1989), 'Likelihood ratio tests for model selection and non-nested hyptheses', *Econometrica* **57**, 307–333.

Wallis, K. F. (2005), 'Combining density and interval forecasts: A modest proposal.', *Oxford Bulletin of Economics and Statistics* **67**, 983–994.

Winkler, R. (1981), 'Combining probability distributions from dependent information sources.', *Managment Science* **27**, 479–488.

A Appendix

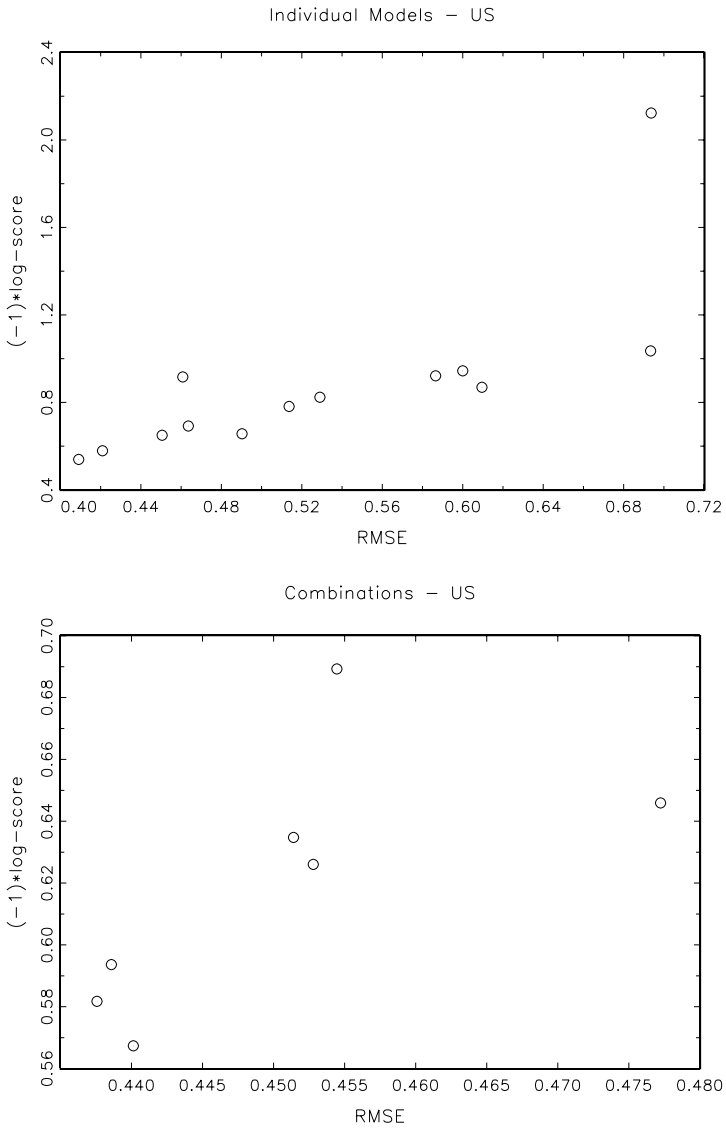

Figure 1: Negative average log-score and RMSPE for the individual and combined density forecasts - US.

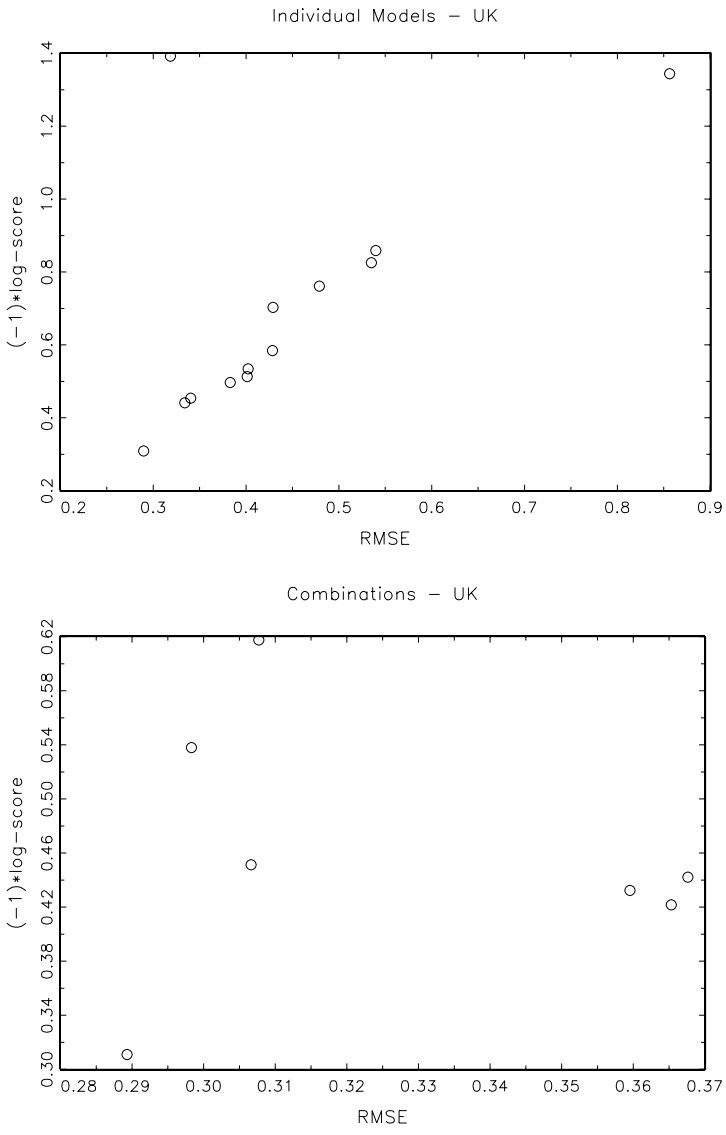

Figure 2: Negative average log-score and RMSPE for the individual and combined density forecasts - UK.

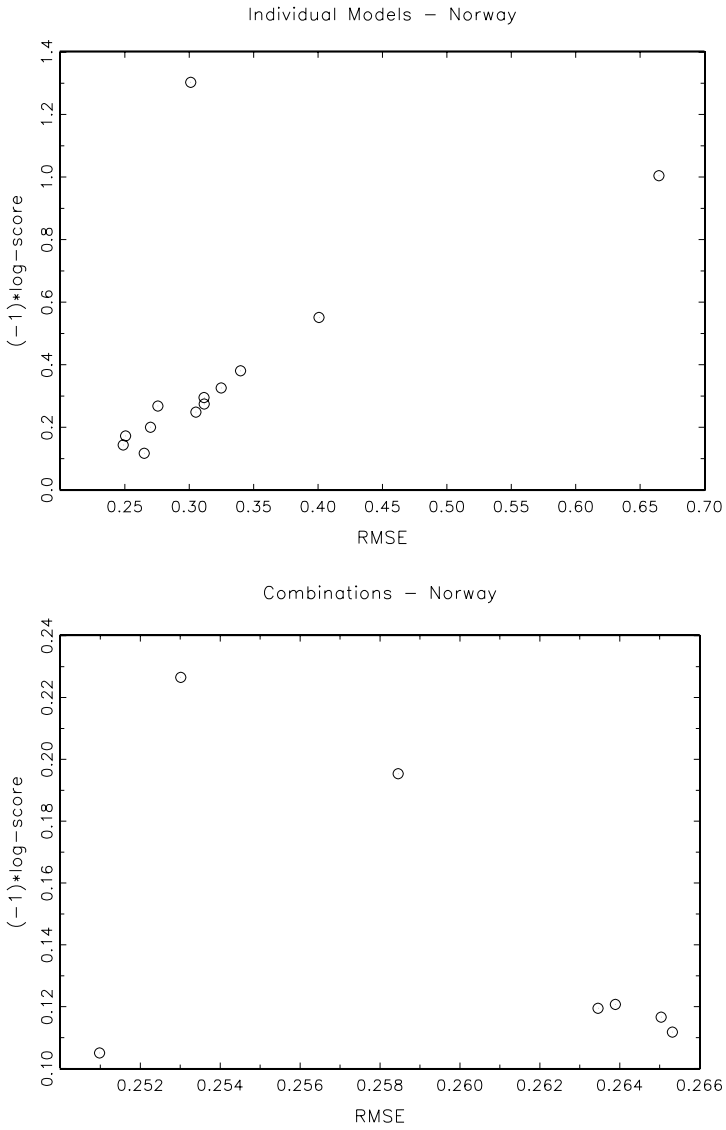

Figure 3: Negative average log-score and RMSPE for the individual and combined density forecasts - Norway.

Table 2: Out-of-sample prediction performance

	US		UK		Norway	
	lnS	RMSPE	lnS	RMSPE	lnS	RMSPE
Individual Models						
RWD4	-0.869	0.610	**-0.309**	0.290	-0.117	0.265
RWD1	-1.035	0.693	-0.703	0.429	-0.275	0.312
RWD1D4	-0.656	0.490	-1.344	0.856	-1.004	0.665
AR1D4	-2.123	0.694	-1.392	0.319	-1.303	0.301
AR1D1	-0.917	0.461	-0.761	0.479	-0.551	0.401
VAR1	-0.650	0.451	-0.859	0.540	-0.381	0.340
VAR2	**-0.539**	**0.409**	-0.825	0.535	-0.326	0.325
VAR3	-0.579	0.421	-0.534	0.402	-0.248	0.305
VAR4	-0.692	0.464	-0.584	0.428	-0.295	0.311
VAR5	-0.922	0.586	-0.497	0.383	-0.172	0.251
VAR6	-0.781	0.514	-0.441	0.334	-0.143	**0.249**
VAR7	-0.824	0.529	-0.453	0.340	-0.200	0.270
VAR8	-0.944	0.600	-0.513	0.401	-0.268	0.276
Selection						
SELEC	-0.626	0.453	-0.442	0.368	-0.117	0.265
Combinations						
EW	-0.689	0.454	-0.618	0.308	-0.227	0.253
RLSW	-0.582	0.438	-0.432	0.360	-0.119	0.263
TRIMW	-0.567	0.440	-0.451	0.307	**-0.105**	0.251
PEW	-0.594	0.439	-0.422	0.365	-0.112	0.265
BMAW	-0.646	0.477	-0.311	**0.289**	-0.121	0.264
BMAPLW	-0.635	0.451	-0.538	0.298	-0.195	0.258

In the table lnS denotes the average log-score evaluated out-of-sample and RMSPE denotes the root mean squared prediction error. See the text for explanation of the model suit.