Asset pricing model comparison incorporating distributional accuracy preferences ☆

Qing Zhou[a,1], Yin Liao[b,*]

[a] *UQ Business School, The University of Queensland, Australia*
[b] *QUT Business School,Economics and Finance, The Queensland University of Technology,Australia*

Abstract

We compare out-of-sample density expectations constructed from 120 empirical specifications of unconditional and conditional asset pricing models. We extend the recently developed model confidence set (MCS) approach to a multivariate setting and use it as a general testing framework to incorporate investors' accuracy preferences for different regions of joint asset return distributions. We find that the performance of unconditional models does not differ significantly from that of conditional models when investors favor the central tendency or the right tail of the density but that the unconditional models outperform the conditional models when investors assess the left tail as more important. We further investigate model pooling as an alternative strategy to address model uncertainty. In contrast to recent studies of asset pricing model pooling, we find that model pooling with a full set of models is not always optimal. We propose a new trimming scheme and find significant gains from trimming prior to pooling, and these gains are most pronounced when investors have a large number of models at their disposal. We provide both statistical and economic explanations for these benefits.

JEL classification: G11, G12, D81, E37

Keywords: Asset pricing model uncertainty, out-of-sample, density

---

[☆]This is a working paper in progress.

[*]Corresponding Author: Yin Liao; Address: QUT Business School,Economics and Finance, The Queensland University of Technology, Australia; Tel: +61 7 3138 2662; Fax: +61 7 3138 1500; Email:yin.liao@qut.edu.au

[1]Email: q.zhou@business.uq.edu.au

1. Introduction

Modeling the relation between risk and expected return is a fascinating endeavor with a long history in finance. To date, the question of whether a particular asset pricing model is literally accurate remains unresolved. As asset pricing models are, at best, approximations of reality, a more useful task for empirical researchers is to compare the performance of competing asset pricing models (Kan and Robotti, 2009). Existing comparison studies are largely based on asset mean return in-sample fitting. However, at least two aspects are important for a comparison of asset pricing models to be ultimately useful for practical applications: the out-of-sample performance (Ferson, Nallareddy, and Xie, 2013) and the reflection of the investor's distributional accuracy preferences, i.e., the preference for the accuracy of different regions of the asset return distribution (Kadan and Liu, 2014). Incorporating these two perspectives, we compare the out-of-sample performance of 120 potential empirical model specifications of individual asset pricing models and three types of asset pricing model pooling in the context of modeling the joint density of cross-sectional stock returns.

We consider all possible empirical specifications of the conditional and unconditional forms of 8 asset pricing models and three sets of widely used instrument variables in prior studies, and we obtain an extensive model set that includes 120 empirical model specifications. By spanning the model space, we aim to address the problem involving asset pricing model uncertainty in a problem-based setting in practical applications in which a large number of models are available to investors. To that end, we first extend the model confidence set (MCS) approach (Hansen, Lunde, and Nason, 2011) to fit our multivariate setting in which the joint density of multiple assets

must be considered. The generosity of the MCS approach also allows us to weight the accuracies of different regions of the predictive asset return distribution such that we can mirror investors' preferences for the distributional accuracy of the center or tail of the asset return distribution in the asset pricing model comparison. In contrast to existing model selection or comparison techniques, the MCS approach acknowledges the limitations of sample data and employs a set of models rather than a single model.

As the sample data commonly used in this area may not be sufficiently informative to identify a single dominant model and as the performance of alternative models also varies over time, model pooling (see Rapach, Strauss, and Zhou, 2010; O'Doherty, Savin, and Tiwari, 2012, for example) has been introduced in recent years to address the uncertainty problem arising in asset pricing models. However, although model comparison or selection risks eliminating useful information that might be contained in the seemingly inferior models, model pooling can also perform poorly when the number of available models is large and contaminating models are included. Because model pooling in finance remains in its early stages with few applications, it is beneficial to the field to emphasize that model pooling cannot be used as a placebo in all conditions to alleviate asset pricing model uncertainty and that pooling all models without concern for the models at hand can be detrimental to investment practices. We propose first trimming the models and then pooling the models using simple equal weights after trimming in the context of out-of-sample asset pricing when empirical researchers have a large number of models at their disposal. In addition to the optimal trimming scheme in the existing literature, we also design a new trimming scheme using the MCS test in the context of asset pricing model uncertainty.

Intuitively, it makes sense to eliminate poor models and pool the remaining models. Our method combining trimming and pooling can be regarded as a tradeoff between model selection and model pooling to address problems involving asset pricing model uncertainty, and this approach has practical value when the model set is large with potentially contaminating models.

Our first main result is that the models selected into the model confidence set vary because of changing preferences for the accuracy in the tail or central tendency along the expected asset return distribution. Consistent with existing studies (e.g., Ghysels, 1998; Simin, 2008), unconditional models generally perform better than their conditional counterparts in terms of numerical comparisons. Previous studies primarily focus on the mean return point forecast comparison, and we complement these studies by showing that unconditional asset pricing models perform better even when investors weigh the central tendency or tail accuracy as more important. However, these conclusions are largely based on numerical comparisons. Our study further tests whether unconditional models perform significantly better than conditional models. In contrast to previous studies, we find that some conditional models cannot be differentiated from unconditional models, particularly when we weigh the central tendency or the right tail as more important. We further show these same findings in two special cases, the point mean and 5% value-at-risk (VaR) loss forecasts, and find that some conditional and some unconditional models are selected into the model confidence set.

This finding provides new insights for the asset pricing literature by showing that the out-of-sample performance of the conditional asset pricing model is close to and cannot be effectively differentiated from that of

unconditional models in forming expectations of tail risk. This result complements the findings in prior asset pricing model comparisons. Recent studies comparing the out-of-sample fit of asset pricing models (e.g., Simin, 2008; Ferson, Nallareddy, and Xie, 2013) primarily focus on the mean point predictive performance of asset pricing models with mean squared error (MSE) and pricing error (PE) used as accuracy measures. However, in real economic applications, investors with different preferences weigh the accuracy of a particular region of the return distribution differently and thus arrive at different rankings of asset pricing models. For example, in risk management, it is critical to examine the tails of the distribution, in which the right tail is important for short positions in the asset and the left tail is critical for downside risk management. The new evidence from our study is of particular interest to this area of study because of the increasing amount of research documenting that asset returns exhibit features in third and fourth moments that are inconsistent with a normal distribution (e.g., Chang, Christoffersen, and Jacobs, 2013; Kelly and Jiang, 2014; Bollerslev, Todorov, and Xu, 2015). Hence, the first two moments, the mean and volatility of asset returns, no longer completely characterize return distributions, and the comparison and evaluation studies of asset pricing models focusing on the point mean and volatility forecasts are less useful than they were under conventional Gaussian assumptions (e.g., Amisano and Giacomini, 2007; Kadan and Liu, 2014). The focus on density forecasting and the incorporation of divergent investor forecasting accuracy preferences distinguish our paper from previous asset pricing model comparison studies, and our study thus has more practical value.

Our second and primary contribution results from introducing trimming

before asset pricing models are pooled to improve the out-of-sample performance of asset pricing models. The pooling approach has long been used in econometric forecasting to improve the out-of-sample performance of economic indicators (see Timmermann, 2006, for a recent review). Finance applications remain scarce; Rapach, Strauss, and Zhou (2010) and O'Doherty, Savin, and Tiwari (2012) are recent attempts in the finance literature to adopt model pooling to improve out-of-sample stock return forecasts. The model sets considered in these papers are relatively condensed, and thus, there is some concern that studies including extremely poor models may actually contaminate the performance of the model pool.[2] However, this consideration is not applicable to real decision making when a large number of potential models are available, and certain poor models are likely to be included. Thus, it is more sensible for investors to dispose of potentially poor models before pooling all models.

In contrast to the common belief in prior studies with pooled asset pricing models, we find that pooling is not always optimal and that trimmed model pools dominate all other models. When the number of available models is large, pooling all models may even underperform the single best model, particularly when extremely poor models are included. Therefore, we propose the trimming of models prior to pooling. After trimming, we find no statistically significant differences between equal- and optimal-weighted model pools. The numerical differences are minor, and trimmed equal weighting generally performs slightly better than optimal weighting. This result is consistent with our intuition that the models in the trimmed

[2]These studies may have applied prior judgment regarding which models to be pooled, but trimming is not explicitly stated in those papers.

model set should perform equally well in statistically meaningful ways if the trimming method is effective. Hence, we suggest pooling models using simple equal weights after trimming. We show that trimming before pooling delivers superior out-of-sample forecasting performance in all cases in the paper. This method provides both statistically and economically significant improvement over pooling with full sets of models examined in prior studies. The result remains the same across different testing assets and across different evaluation periods. Moreover, our proposed method is easy to implement in practice, and from a practical perspective, it is meaningful to discard poor models when model spaces are large and extreme models are potentially incorporated.

In addition to our main contributions to the asset pricing model uncertainty literature, we also contribute to the forecast combination literature by proposing a new method to trim models, and we provide evidence of the advantages of trimming in the context of out-of-sample density forecasting of asset pricing models, which involves numerous potential empirical model specifications. We are the first researchers in the field to specifically create MCS tests for asset pricing model comparisons. To the best of our knowledge, we are also the first in this field of study to span empirical asset pricing model specifications to such a large model set that includes a relatively comprehensive grouping of potential empirical model specifications.

The remainder of this paper is organized as follows. Section 2 describes the asset pricing models and various model specifications considered in our study and provides an overview of previous asset pricing model comparison studies. Section 3 presents the empirical methods utilized in our study and discusses empirical design issues. Section 4 illustrates the construction of

the data sample. Section 5 discusses the empirical results, and Section 6 provides statistical and economic interpretations of the empirical results. Section 7 addresses other empirical design issues and presents the results of robustness tests, and Section 8 concludes the paper.

2. Asset Pricing Models

The models that we examine are derived from the asset pricing literature. We begin with the most celebrated asset pricing model, the capital asset pricing model (CAPM), and proceed to the advancement of Merton's 1973 intertemporal capital asset pricing model, for example, and the consumption CAPM (CCAPM) (Breeden, 1979), the Chen, Roll, and Ross 1986 5-factor model (CCR5) and Jagannathan and Wang's 1996 conditional version of the CAPM (JW). We then proceed to the multi-factor models of Fama and French (1993), Carhart (1997) and Liu (2006), which are motivated by the empirical failure of the CAPM and return anomalies. These models are widely used in studies that compare and evaluate asset pricing models (e.g., Hodrick and Zhang, 2001; Simin, 2008; Kan and Robotti, 2009; O'Doherty, Savin, and Tiwari, 2012). Our primary model set thus comprises eight asset pricing models: (1) the Sharpe (1964)-Lintner (1965)-Mossin (1966) CAPM; (2) the linear version of the CCAPM; (3) the CCR5; (4) the JW; (5) the Fama and French (1993) 3-factor pricing model (FF3); (6) the Carhart (1997) 4-factor model (Carhart4); (7) the Fama and French (2015) 5-factor pricing model (FF5); and (8) the liquidity-augmented asset pricing model (LIQ) by Liu (2006).

The general empirical specification for each of these models is a multivariate normal linear regression with random regressors (O'Doherty, Savin,

and Tiwari, 2012)

$$r_t = \alpha_t + \beta_t f_t + \epsilon_t \tag{2.1}$$

where r_t is an $m \times 1$ vector of excess returns at time t for a set of m assets, f_t is a $k \times 1$ vector of k factors at time t, and ϵ_t is an m vector of disturbances at time t that is normally independent and identically distributed (IID) with a mean of 0 and a positive definite variance matrix Σ. The true specification of the models does not contain an intercept term, i.e., $\alpha_t = 0$.

To include a relatively comprehensive set of asset pricing models, we consider various empirical variations of the asset pricing models, including both unconditional and conditional models. We also allow for time-varying risk premiums and time variations in the intercept and the beta. The various specifications of equation(2.1) are as follows:

$$r_t = \beta_0 f_t + \epsilon_t, \tag{2.2}$$

$$r_t = \alpha_0 + \beta_0 f_t + \epsilon_t, \tag{2.3}$$

$$r_t = \alpha_0 + (\beta_0 + \beta_1^{'} Z_{t-1}) f_t + \epsilon_t, \tag{2.4}$$

$$r_t = (\alpha_0 + \alpha_1^{'} Z_{t-1}) + \beta_0 f_t + \epsilon_t, \tag{2.5}$$

$$r_t = (\alpha_0 + \alpha_1^{'} Z_{t-1}) + (\beta_0 + \beta_1^{'} Z_{t-1}) f_t + \epsilon_t \tag{2.6}$$

$$r_t = \beta_0 (C_0 + C_1^{'} Z_{t-1}) + \epsilon_t, \tag{2.7}$$

$$r_t = \alpha_0 + \beta_0 (C_0 + C_1^{'} Z_{t-1}) + \epsilon_t, \tag{2.8}$$

$$r_t = (\beta_0 + \beta_1^{'} Z_{t-1})' (C_0 + C_1^{'} Z_{t-1}) + \epsilon_t, \tag{2.9}$$

$$r_t = (\alpha_0 + \alpha_1^{'} Z_{t-1}) + (\beta_0 + \beta_1^{'} Z_{t-1})' (C_0 + C_1^{'} Z_{t-1}) + \epsilon_t. \tag{2.10}$$

where z_{t-1} is the vector of lagged instruments. Equation 2.2 and 2.3 represent an unconditional specification of the models with a constant intercept,

beta and factor risk premium. The remaining equations are conditional specifications with time variation in the coefficients and/or risk premiums. These conditional specifications allow us to examine whether time variation in the intercept, beta and factor risk premiums improves the out-of-sample fit of the models. Similar specifications can also be found in Simin (2008). To estimate the specifications (2.4), (2.5) and (2.6) with time variation in the intercept and/or beta, we use three different sets of instrumental variables to construct a lagged information set, Z_{t-1}, which is detailed in the data section. Model specifications (2.7) to (2.10) involve time variation in the factor risk premium. To avoid perfect multicollinearity in estimating multi-factor conditional models, we select different instruments from the six variables to determine the factor risk premium of each distinct pricing factor.

The modeling freedom with respect to the model form and instrumental variables lead to a total of 120 empirical model specifications, i.e., 15 specifications for each of the eight asset pricing models, which includes two unconditional models, nine conditional specifications incorporating time variation in coefficients (three model forms with three different vectors of instruments) and four conditional specifications with time variation in factor risk premiums. To simplify references to the models in the following sections, we include notations for each of the 120 specifications in AppendixA and index the models from 1 to 120.

3. Empirical Methods

3.1. Constructing a density forecast from the asset pricing models

We illustrate our method to obtain one-step-ahead [3] asset return density forecasts in this subsection. The objective of a density forecast is to obtain the forecast of full predictive density of asset returns rather than merely forecasts that match the observed data in the first or second moments, for instance. In the setting of cross-sectional asset pricing, obtaining predictive densities for multiple assets simultaneously is of greater interest in asset allocation practice. We form the density forecasts entirely in the multivariate setting.

We implement a Bayesian approach to obtain the predictive density of r_t conditional on the set of factors f_t and the information set Z_{t-1}. Assume that the investor has the standard uninformative prior on B:

$$p(B, \Sigma) \propto |\Sigma|^{\frac{m+1}{2}}. \tag{3.1}$$

Using similar specifications as in O'Doherty, Savin, and Tiwari (2012), we let the marginal posterior probability density function (PDF) for Σ follow an inverse Wishart distribution and the conditional posterior PDF for B follow a multivariate normal distribution. The expression for the one-step-ahead conditional predictive density for r_t takes the form of the multivariate

[3] The forecast horizon is another a crucial object in measuring and ranking forecast accuracy (Diebold and Lopez, 1996). We limit our treatment to the one-step-ahead forecast to concentrate on our main points and leave investigations of this point for future studies.

Student's t-distribution:

$$
\begin{aligned}
& p(r_t|R_{t-\tau,t-1}, F_{t-\tau,t-1}, f_t, Z_{t-1}) \\
& = \frac{v^{1/2}\Gamma[(v+m)/2]|V|^{1/2}}{\pi^{m/2}\Gamma(v/2)} \\
& \times [v + (r_t - \hat{\alpha}_t - \hat{\beta}_t f_t)' V (r_t - \hat{\alpha}_t - \hat{\beta}_t f_t)]^{-\frac{v+m}{2}}, \qquad (3.2)
\end{aligned}
$$

where τ is the length of the sample and the subscript $t-\tau, t-1$ denotes a sample that extends from time $t-\tau$ to $t-1$. The ith row of the $\tau \times m$ matrix $R_{t-\tau,t-1} = r_{t-\tau}, ..., r_{t-1}$ is $r_{t-\tau-1+i}$, the ith row of the $\tau \times (k+1)$ matrix $F_{t-\tau,t-1}$ is $(1, f'_{t-\tau-1+i})$, and f_t is the vector of 1-step-ahead factor realizations at time t. $V = gvS^{-1}$, $g = 1 - f'_t(\overline{F}'\overline{F} + f'_t f_t)^{-1} f_t$, $\overline{F} = \{f_{t-\tau}, ..., f_{t-1}\}$, and $v = \tau - (k+1) - (m-1)$.

Notably, our objective here is the one-step-ahead predictive density of r_t conditional on f_t rather than a marginal distribution of r_t. As shown in O'Doherty, Savin, and Tiwari (2012), under the assumption that the asset pricing models share a common prediction model for factor returns, the marginal distribution of r_t can be obtained by integrating out f_t from the joint distribution of r_t and f_t, which is therefore reduced to a conditional distribution of r_t on f_t. Hence, the evaluation of the asset pricing models does not require the explicit specification of a prediction model for the factors, and the density in equation (3.2) can actually be used to evaluate the true out-of-sample predictive performance for a set of competing models.

To assess the performance of alternative density forecasting models and methods, we use a likelihood-based metric, i.e., the log predictive score (LPS) function, which is closely related to the Kullback-Leibler distance.

Denoting the realizations of time series r_t as r_t^o, the LPS function of the joint predictive density of r_t is

$$LS = \sum_{t=2}^{T} log[p(r_t^o, f_t^o | R_{t-\tau,t-1}, F_{t-\tau,t-1}, f_t, Z_{t-1})]. \tag{3.3}$$

The LPS function is intuitively appealing because it reflects the out-of-sample prediction performance of a given model throughout while giving a high score to a model that ex ante assigns a high probability to the value of the r_t that materializes. Therefore, a model that maximizes the LPS function is equivalent to those that minimize the distance between the forecast density and the true but unknown density measured by the Kullback-Leibler information criterion (KLIC).

3.2. Comparing the out-of-sample performance of alternative models

We use the MCS approach to compare alternative asset pricing models and forecasting methods. Instead of making comparisons solely based on absolute numbers, we can use an MCS test to make statistical inferences regarding the significance of differences across models. This method was initially developed by Hansen, Lunde, and Nason (2011), and we extend the method to cross-sectional asset pricing model density forecast comparison and incorporate various preferences for the accuracy of particular regions of the return distribution. This method has two advantages. First, by acknowledging possible limitations in the data because the number of models chosen containing the best model will depend on how informative the data are, the proposed method selects a model set containing the best-performing models at the confidence level of α, with α representing the size of the MCS test rather than selecting a single-best model. Therefore, the MCS approach

features the properties of both model selection and model trimming. Second, as the ranking of density forecasts is rarely independent of individual preferences (Diebold, Gunther, and Tay, 1998), particularly for various financial applications, such as portfolio selection, risk management, utility or objective functions varying across decision makers, the loss functions must be generated specifically for different preferences for the accuracies of distributional forecasts. The MCS is a flexible method that can be used for arbitrary loss functions. Therefore, we can use a weighting function such as the functions proposed by Amisano and Giacomini (2007) to assign weights to different parts of the distribution. For example, using the normal density function will give more weight to the center of the distribution, and one minus the normal cumulative distribution function (CDF) will allocate more weight to the left tail of the distribution, whereas using the normal CDF will result in more weight given to the right tail. The forecasting accuracy of tails is critical to risk management. In this subsection, we first describe the general testing framework and then use a weighting function to assign weights to the forecasting losses of the different regions of asset return distributions the enable the incorporation of investor distributional accuracy preferences. We further provide two special cases of different accuracy preferences: the point mean return and VaR forecasts.

3.2.1. A general testing framework: the model confidence set approach

We define a finite model set $\mathcal{M}_{\prime}$ containing N asset pricing models, indicating the initial number of asset pricing models in our empirical procedure. The MCS approach (see Hansen, Lunde, and Nason, 2011, for a more

detailed exposition) aims to identify the set $\mathcal{M}^*$ such that

$$M* = \{i \in M_0 : u_{i,j} \leq 0 \, for \, all \, j \in M_0\}, \tag{3.4}$$

where $u_{i,j} = E(d_{ij,t})$ is the expected loss differential between models and $d_{ij,t}$ represents the loss differential between models i and j that is defined as

$$d_{ij,t} = L_{i,t} - L_{j,t}, \qquad \text{for all} \, i, \, j \in M_0.. \tag{3.5}$$

where L_t denotes the loss associated with object i or j. This loss function can also be customized for an investor by introducing a weighting function, which yields a more general relative performance measure that is described as follows:

$$\hat{d}_{ij,t} = w(r_t)d_{ij,t}, \qquad t = 1, ..., T. \tag{3.6}$$

In other words, given the set of all forecasting models $\mathcal{M}$, in the comparison set, the MCS searches for the set of models that cannot be rejected as statistically inferior at a chosen level of confidence.

The implementation of MCS is based on the following algorithm: beginning with the set of all models $\mathcal{M}$, at significance level α, the null hypothesis

$$H_0 \; : \; \mathrm{E}(\hat{d}_{ij,t}) = 0, \qquad \forall \; i > j \in \mathcal{M}, \tag{3.7}$$

is tested. If H_0 is rejected at the significance level α, then the worst-performing model is removed, and the process continues until non-rejection occurs with the set of surviving models being the MCS, $\widehat{\mathcal{M}}^*_\alpha$. If a fixed significance level α is used at each step, then $\widehat{\mathcal{M}}^*_\alpha$ contains the best model

from $\mathcal{M}_0$, with $(1-\alpha)$ as the level of confidence[4].

The relevant t-statistic, t_{ij}, provides scaled information on the average difference in the forecast quality of models i and j and is defined as

$$t_{ij} = \frac{\overline{d}_{ij}}{\sqrt{\widehat{\text{var}}(\overline{d}_{ij})}}, \qquad \overline{d}_{ij} = \frac{1}{T}\sum_{t=1}^{T} d_{ij,t}\,. \tag{3.8}$$

where $\widehat{\text{var}}(\overline{d}_{ij})$ is an estimate of $\text{var}(\overline{d}_{ij})$ and is obtained from a bootstrap procedure described in Hansen, Lunde, and Nason (2003). In this paper, the null hypothesis in equation ((3.7)) is tested using the range statistic

$$T_R = \max_{i,j\in\mathcal{M}} |t_{ij}| \tag{3.9}$$

with all p-values obtained by bootstrapping the necessary values 1,000 times. When the null hypothesis is rejected, the worst-performing model is identified as

$$i = \arg\max_{i\in\mathcal{M}} \frac{\overline{d}_{i}}{\sqrt{\widehat{\text{var}}(\overline{d}_{i.})}}, \qquad \overline{d}_{i.} = \frac{1}{m-1}\sum_{j\in\mathcal{M}} \overline{d}_{ij}\,. \tag{3.10}$$

The MCS test is a tool to summarize the relative performance of an entire set of models by determining which models can be considered statistically superior and at what level of significance. The interpretation of an MCS for a set of models is analogous to a confidence interval for a parameter in the sense that the models covered by the MCS are those that cannot be rejected from the set of best models for a given level of confidence. By attaching p-values to models, we can easily determine the level of significance at which individual models can be in the set of superior models and which can be

[4]Although the testing procedure involves multiple hypothesis tests, this interpretation is statistically accurate. See Hansen, Lunde, and Nason (2003) for details.

eliminated as statistically inferior.

3.2.2. Incorporating accuracy preferences

Most practical applications involving joint density forecasts of multivariate asset returns focus on a particular region of the expected distribution of asset returns. For example, in risk management, the main concern for banks and other financial institutions is to accurately describe the left tail of the distribution of their portfolio returns to obtain accurate estimates of downside risk (e.g., VaR). Thus, incorporating the decision maker's or the investor's distributional accuracy preferences is important.

Because a density expectation can be understood as a collection of probabilities assigned by investors to all attainable events, we use the weighted scoring rules of Amisano and Giacomini (2007) and adapt it to multivariate density forecasts. These rules allow investors to assign greater weight to particular regions of the distribution of the asset return. In some cases, the investor may be particularly interested in predicting "tail events," as these can lead to different asset allocation decisions, whereas in other scenarios, the investor is more interested in knowing future events that may fall near the center of the distribution and ignores the influence of possible outliers. We can define an appropriate weight function $w(.)$ and compare the weighted average scores

$$n^{-1}\sum_{t=2}^{T} w(r_t) logp(r_t|R_{t-\tau,t-1}, F_{t-\tau,t-1}, f_t, Z_{t-1}). \tag{3.11}$$

The weighting function $w(.)$ can be chosen by the investor to select the desired region of the distribution of r_t. As the predicted joint return follows a multivariate Student's t-distribution, we can use the following weights for

different regions of the density:

- Central tendency: $w_1(r_t) = mvtpdf(r_t)$, where $mvtpdf(r_t)$ is the standard density function (PDF) of the multivariate t-distribution.
- Right tail: $w_2(r_t) = mvtcdf(r_t)$, where $mvtcdf(r_t)$ is the standard probability function (CDF) of the multivariate t-distribution.
- Left tail: $w_3(r_t) = 1 - mvtcdf(r_t)$, where $mvtcdf(r_t)$ is the standard probability function (CDF) of the multivariate t-distribution.

The weighted scoring rule can be customized for the MCS test by using a more general relative performance measure, as stated in equation (3.6).

3.2.3. Two special cases of accuracy preferences

Studies in the asset pricing literature commonly compare the point forecasts of mean asset returns (e.g., Simin, 2008; Ferson, Nallareddy, and Xie, 2013). The assumption underlying this choice is that investors care more about the central tendency of the predicted return distribution. We include point mean return forecasting as a special case that places more weight on the accuracy of the center of the predicted distribution.

Following Fama and French (1996), we evaluate the performance of point mean return forecasting using the average model pricing error. For a given portfolio, the pricing error (PE) is calculated as follows:

$$PE = \frac{1}{T}\sum_{t=1}^{T}(r_t - \hat{r}_t) \tag{3.12}$$

The PE is the time-series average of the monthly difference between the realized and predicted portfolio returns, where r_t is the realized return and

$\hat{r}_t$ is the return forecast from alternative forecasting methods. We use two measures based on PE: one is cross-sectional average absolute pricing error across portfolios (Ave $|PE|$), and the other is the average squared pricing error across assets (Ave PE^2). These two measures are widely used in the prior literature to evaluate the performance of asset pricing models (e.g., Simin, 2008).

A second special case is the assessment of the downside risk faced by an investor. We examine the alternative models in forming expectations regarding 5% VaR, which is the 5% quantile of the predicted asset return distribution and can be understood as an example of interval forecasting that emphasizes the tail of the distribution. It can succinctly convey the uncertainty surrounding a point forecast. A 5% VaR forecast can be interpreted as an interval forecast for future returns, with an interval of $(-\infty, VaR]$ and a confidence level of 95%.

To forecast the tail interval of the joint density of the cross-sectional return r_t, we form a portfolio of the set of risky assets. Under the assumption that the predictive density of r_t has the form of a multivariate Student's t-distribution, the VaR forecast of this portfolio can be obtained as follows:

$$VaR^{\tau}_{t|F_{t-\tau,t-1}} = \mu_{p,t} - \sqrt{\frac{\nu-2}{\nu}}\sigma_{p,t}t_{1-\tau,\nu}, \tag{3.13}$$

where $\mu_{p,t} = \omega_t E(r_t)$ and $\sigma_{p,t} = \sqrt{\omega_t \Sigma_t \omega_t'}$. $E(r_t)$ denotes the $m \times 1$ vector of expected excess returns on risky assets at time t, and Σ_t is the corresponding covariance matrix estimate based on the predictive return distribution. ω_t represents the weight of each risky asset in the portfolio. We consider only equally weighted portfolios for simplicity, but the arguments below hold

with equal force for any well-diversified portfolio.

To compare the out-of-sample interval forecast, we consider the asymmetric VaR loss function of González-Rivera, Lee, and Mishra (2004) to compare the ability of different asset pricing models to predict extreme loss in returns. The asymmetric VaR loss function of González-Rivera, Lee, and Mishra (2004) is defined as

$$l(r_t, VaR^{\tau}_{t|F_{t-\tau,t-1}}) = (\tau - I^{\tau}_t)(r_t - VaR^{\tau}_{t|F_{t-\tau,t-1}}), \tag{3.14}$$

where r_t is the portfolio return at time t, $VaR_{t|F_{t-\tau,t-1}}$ denotes the τ level predicted VaR at time t conditional on the information set $F_{t-\tau,t-1}$, and $I_t = 1(r_t < VaR_t)$ is an indicator function that equals 1 when r_t is below VaR_t and 0 otherwise. The asymmetric VaR loss function represents the natural candidate to evaluate the asset pricing model performance for tail risk prediction because it more severely penalizes observations below the τth quantile level.

3.3. Asset pricing model trimming and pooling

In forming the expectations of asset return density in a model-rich environment with model uncertainty, two general approaches are commonly used: model selection and model pooling. Model pooling has often been used to improve upon single-model forecasts. As noted in the review by Timmermann (2006), combining models can diversify model uncertainties and alleviate the structural breaks and misspecifications plaguing individual models. The model pooling approaches are in the spirit of Box (1980), i.e., all models are false, but some are useful. However, in the context of a large set of potential models, without prior knowledge of the individual models,

inclusion of a full set of models can be problematic because only a fraction of the forecasting models is likely to contain valuable information regarding the target variable, while the remainder of the models are pure noise (Timmermann, 2006). Moreover, when combination weights must be estimated, the estimation error induced by additional parameters can outweigh the marginal diversification gains from the additional poor models, particularly when the full model set may contain implausible predictions. Hence, prior to combining models, the more important question is how to find "useful models" to avoid introducing extremely poor models that can contaminate the pooled forecast. We advocate that it is advantageous to trim models before pooling them. This approach echoes the sentiment of Armstrong (1989) to "use sensible models." In this subsection, we first discuss the out-of-sample density forecast from the asset pricing model selection and then describe the trimming and pooling method.

3.3.1. Trimming before pooling

One of the common trimming schemes is optimal fixed trimming. In a fixed trimming scheme, the number of forecasting models to be discarded is determined exogenously. The normal practice of fixed trimming is to rank models according to certain evaluation metrics, to discard a fixed proportion of models and to use the remaining models to generate the set of best forecasts. We determine the optimal trimming percentage by maximizing the LPS of the density forecast from the trimmed model pool. As the number of models to be discarded (and hence pooled) is exogenously fixed, discarding the same models for each forecast period is unnecessary. Hence, in our empirical implementation, we trim and include different sets of models

based on the applicable model rank in the periods preceding the out-of-sample testing period.

We propose a new trimming scheme based on the MCS test. Instead of selecting a single best model, the MCS test can select a set of best models that perform the same statistically, and the set of models that perform statistically worse can be eliminated. Hence, the MCS approach provides a natural way to trim models before pooling. When constructing the best model set, we begin from the full set of the model candidates and conduct the MCS test sequentially at a confidence level of 0.1% to retain models with an associated p-value greater than or equal to 0.1%.

Optimal fixed trimming is as simple as a grid-search method; this trimming method begins with an initial percentage to be trimmed in an attempt to reveal a trimming percentage that forms a model pool to achieve the best forecast. Optimal fixed trimming is fully free of statistical test distortions and can be easily implemented. An important difference between the two schemes is that MCS trimming accounts for the statistical significance of differences in the historical performance of the models, but optimal fixed trimming does not.

We generate the out-of-sample expected density by aggregating the density forecasts from each model contained in the selected model set. The aggregation uses both optimally and equally weighted averages of the model forecasts. We provide details of the weighting in the following subsection.

3.3.2. Model pooling

The key input for a model pool is the weight assigned to individual models. Generally, the combination weights can be estimated by solving,

minimizing or maximizing an objective function. We obtain the model weights for optimal density combination by maximizing the log predictive score criterion, which incorporates the predictive densities of each selected model.

We adopt the optimal pooling scheme as shown in O'Doherty, Savin, and Tiwari (2012). The model weights w are chosen to maximize the LS function:

$$f_T(w) = \sum_{t=2}^{T} \sum_{i=1}^{N} w_i p_i(r_t | R_{t-\tau,t-1}, F_{t-\tau,t-1}, f_t, Z_{t-1}), \tag{3.15}$$

subject to the restrictions $\sum_{i=1}^{N} w_i = 1$ and $w_i \geq 0$. $p_i(r_t | R_{t-\tau,t-1}, F_{t-\tau,t-1}, f_t, Z_{t-1})$ is the predictive density implied by the ith model. Accordingly, the optimal prediction pool corresponds to

$$w^* = arg \max_{w} f_T(w). \tag{3.16}$$

We also consider the simple-average weighting scheme that assigns equal weights to alternative single models, i.e., $w_{i,t} = 1/N$ for the ith model, where $i = 1, ..., N$. The simple "1/N" rule is found to be dominant in more refined combination schemes in many empirical forecast combination studies (e.g., Timmermann, 2006; Smith and Wallis, 2009).

3.4. Empirical Design Issues

3.4.1. Empirical implementation

Following standard practice, we estimate all model parameters using a five-year rolling window of monthly time-series data (i.e., $\tau = 60$). Therefore, the first density forecast is generated for time $t = 61$ (i.e., July 1968)

using each of the individual asset pricing models.

For a pool of models, the optimal weights are calculated by maximizing the LPSs of the density forecast, as shown in equation (3.3). The model weights are obtained from three designs: the full sample, rolling 10-year windows, and an expanding window. In the case of the full sample, the LPS is calculated for time T using monthly predictive densities from time $t = 61$ to time $t = T$, where T is the last month of the sample data (i.e., December 2011). In the case of the rolling window, the model weights are estimated using a rolling window of 10 years (i.e., 120 months); thus, for the rolling window, the first pooled forecast is made for July 1978 using the data from July 1968 to June 1978. For the expanding window, the first LPS is still calculated using predictive densities based on an initial 10-year window from July 1968 to June 1978, and subsequent LPSs are updated monthly based on an expanding window of data. In each case, the model weights are determined at time t using only the information available through time $t - 1$.

3.4.2. Sample split and subperiods

Our main focus is on the out-of-sample performance of the alternative asset pricing models. For the out-of-sample forecast, one important choice is the split point of the in-sample estimation and the out-of-sample evaluation periods. As noted by Hansen and Timmermann (2012), out-of-sample tests rely on how a given data set is partitioned into estimation and evaluation subsamples. However, to date, there has been no guidance or consensus regarding how to choose the split point. As the degree of power is the highest when the forecast evaluation window begins early, corresponding to

a long out-of-sample period (Hansen and Timmermann, 2012), we retain the estimation sample in our main results to include 60 monthly observations, i.e., five years as opposed to the 521 monthly observations in the evaluation period. To explore the robustness of our results, we choose different lengths of estimation periods. We use 120 and 300 monthly observations and report the results in the robustness check section.

In addition to the sample split point, another concern may arise from the robustness of our results in distinct subperiods. Hence, we also investigate whether our results vary in several subsamples. We divide our testing sample into five subperiods, including periods prior to and following the 2008 global financial crisis.

4. Data

Our sample contains three sets of data: 1) a return series of testing assets, 2) a series of pricing factors, and 3) a series of instrumental variables. Our sample period spans from July 1963 to December 2011. We use both the 30 value-weighted industry portfolios and the 25 size and book-to-market (B/M) portfolios as our testing assets. We select these assets to alleviate concerns regarding the sensitivity of asset pricing testing to the selected testing assets(see Lewellen, Nagel, and Shanken, 2010; Kan, Robotti, and Shanken, 2013). The monthly portfolio return data are retrieved from Kenneth French's website. The monthly excess asset return is calculated as the gross return in excess of the one-month T-bill rate obtained from Ibbotson Associates.

Our base model set includes eight asset pricing models that are detailed in Section Pricing models. The data source and construction of the pricing

factors of alternative models are provided in AppendixA. We use six commonly used instrumental variables for specifying the conditional models. Following Kan and Robotti (2009) and Hodrick and Zhang (2001), we use the cyclical part of the natural logarithm of the industrial production index lagged one period (Lag IP) and a January dummy (JAN) as two separate instruments. We follow Simin (2008) to group four variables into a vector of instrument variables: the U.S. one-month T-bill from the CRSP Risk-Free Rates file (T-bill), the dividend yield of Standard & Poor's 500 index (DivYld), the spread between the lagged Moody's Composite Average of Yields on Corporate Bonds and the U.S. one-month T-bill from the CRSP Risk-Free Rates file (Term), and the difference between the Moody's BAA and AAA corporate bond yield (Junk). Hence, we obtain three sets of instrument variables, including two univariate and one multivariate information sets.

5. Empirical Results

We summarize the main results of our study in the two following subsections, which focus on two asset pricing model comparisons and the merits of asset pricing model pooling, respectively.

5.1. Asset pricing model comparisons

The out-of-sample density forecasting performance of the asset pricing models will first be compared using LPS with different accuracy preferences. Table 1 reports the summary statistics for (weighted) LPSs for each individual model using the full sample. The 2nd and 3rd columns of the table present results for 30 industry portfolios, with Panel A, B, C and D

respectively reporting results for entire region, the central tendency and the left/right tails of the density. Over the entire sample period, the Fama and French (2015) five-factor model with constant beta performs the best for the entire region of the density and its left tail, as evidenced by the model's (weighted) log score function values of -58806.5269 and -55728.2537, and the Jagannathan and Wang (1996) CAPM model and the Fama and French (1993) three-factor model with a constant beta perform the best in the case of the central tendency and the right tail of the density, with (weighted) log score function values of -5.9199E-25 and -3063.0472, respectively, which is the highest of the 120 competing models. By contrast, the Chen, Roll, and Ross (1986) five-factor model with a time-varying intercept and beta using the third set of instruments perform the worst in all cases.

For comparison, the 4th and 5th columns of Table 1 report results using 25 size and B/M portfolios as test assets. As might be expected, the Fama-French (1993) model with a constant beta should outperform the other models because this model includes factors particularly designed to "explain" returns on a cross-section of portfolios sorted by size and B/M equity (O'Doherty, Savin, and Tiwari, 2012). Consistent with our expectations, the Fama-French (1993) model does exhibit the best out-of-sample predictive ability in most cases, except for the central tendency, as measured by the model's (weighted) log score function values of -28337.6825, -22514.7137 and -5822.9689. The Liu (2006) liquidity-augmented asset pricing model with a time-varying beta using the third set of instruments performs the best for the central part of the density, with a weighted log score function value of -1.0103E-12. Similarly, the worst-performing model is the Chen, Roll, and Ross (1986) five-factor model. The specification with the

time-varying intercept, beta and risk premium is the worst for the central tendency, and the specification with the time-varying intercept and beta is the worst for the other cases.

[Insert Table 1 here.]

Although the aforementioned results indicate the relative accuracy of forecasts, they do not indicate whether any differences in performance are significant. The MCS test of Hansen, Lunde, and Nason (2011) will be used to achieve determine this, and the results are displayed in Figure 1. The dots in the plot present the (weighted) LPS of the density forecast from the corresponding model, and the straight red line in the plot indicates the threshold above which the models are selected into the best model set based on the MCS test, which is conducted at a significance level of 0.1%. When equal weights are used to evaluate the density forecast, FF5(Ub), Carhart4(Ub) and FF3(Ub) are the models selected into the best model set with a p-value greater than 0.001, and this result implies that those three models are significantly better then all the others at a significant level.When focusing only on the central region of the density, all models show some similarity and are included in the best model set. When shifting attention to the tails of the density with desirable application in risk management, Carhart4(Ub)and FF5(Ub) are two models selected into the superior set for the left tail, and the majority of models are selected into the superior set for the right tail, with the main exceptions being the CCAPM models. For the case using 25 size and B/M portfolios as test assets, the MCS test is also implemented at the significance level of 0.1%. Carhar4(Ub), Carhart4(Uab), FF3(Ub), FF3(Uab) and FF5(Ub) outperform all the other models and are selected into the superior set when equal weights are used. Similar to what

we observed in 30 industry portfolios, all models have p-values greater than in terms of the central part of the density forecast, and they are all selected to form the best model set. With respect to the tail forecast, while the left tail continues to require an aggressive eliminating scheme by including only Carhart4(Ub), FF3(Ub), FF3(Uab) and FF5(Ub) in the best model set, the right tail also appears less generous than that in the 30 industry portfolios, given that only Carhart4(Ub), Carhart4(Uab), Carhart4(Ca1Ub), FF3(Ub), FF3(Uab), FF5(Ub), FF5(Uab) and FF5(Ca1Ub) are included in the best model set.

[Insert Figure 1 here.]

In summary, these results reveal several noteworthy points. First, the Fama and French (1993) 3-factor pricing model, the Carhart (1997) 4-factor model, and the Fama and French (2015) 5-factor pricing model outperform the other models in describing both the center and tail of the distribution. These three models are most frequently selected regardless of which region of the density on which the investor is more focused. Second, the conditional asset pricing models are found to be useful in density forecasts of the return and particularly for the tail forecast, with superior performance to many unconditional models. This finding contrasts with the previous literature that primarily focuses on the mean point predictive performance of asset pricing models and claims that the conditional asset pricing models perform much worse than unconditional models. The finding of their important roles in tail forecasting is particularly interesting, as it implies that the conditional models may be desirable in risk management. Finally, the tail forecast of the density is more selective for models than the central tendency, and accuracy in its forecasting is more difficult to achieve. Although all asset

pricing models are selected into the best model set for central tendency forecasting of the density, far fewer models are included in the best model set for tail forecasting.

These results are further corroborated by two special cases. Table 2 and Table 3 present results for both center and tail point forecasts that focus only on the mean and 5% VaR of the distribution.[5] Consistent with the density forecast, the conditional models are found to be useful for both center and tail point forecasts. Although the conditional CCAPM model is included in the best model set for mean point forecasts of the 30 industry portfolios based on Ave $|PE|$, both the conditional CCAPM model and the conditional Fama and French (1993) 3-factor model are selected into the best model set for the mean point forecast of 25 size and B/M portfolios based on both Ave $|PE|$ and Ave PE^2 at the 5% significance level. In tail point forecasting, most conditional models are included in the superior model set for 30 industry portfolios, and conditional CRR5 is selected for 25 size and B/M portfolios at the 5% significance level.

[Insert Table 2 here.]

[Insert Table 3 here.]

5.2. The comparison of alternative strategies in addressing model uncertainty

In this section, we consider four forecasting schemes using the 120 asset pricing models to compare their out-of-sample forecasting performance in terms of the entire part of the return distribution, its central tendency, and

[5]Because of space limitations, we include only representative results of a proportion of the models here, and the results for the full set of models are available upon request.

the left and right tails. The forecasting schemes considered here include the best individual model, the model pooling with the full set of models, and the model pooling with MCS and optimal trimming.

Table 4 shows the forecasting results for the full density. Columns 2 to 4 report the results based on the full sample, rolling samples and expanding samples. Panels A and B respectively present results for 30 industry and 25 size and B/M portfolios, respectively. First, we find that combining models in the best model set constructed using MCS tests outperforms the best individual model in all cases, which verifies that all individual asset models may be subject to misspecification bias of an unknown form (Kan and Robotti, 2009). When the data are not adequately informative, identifying a single dominant model is not possible. The MCS test approach can be treated as another model selection device; it selects the set of best models with the level of confidence that one selects and implicitly trims models that are not in this set. An advantage of the test is that it does not necessarily select a single model; it instead acknowledges possible limitations in the data. Given a certain confidence level, if the data are not informative, many models will be included in the set of best models, as the MCS will have difficulty distinguishing between the models, and few will be trimmed. By contrast, if the data are informative, the set of best models will consist of fewer models. Meanwhile, the performance of the equal weighting scheme is similar to that of the optimal weight scheme when combining models in the best model set. This result is not surprising and confirms our intuition that models included in the best model set are not distinguishable from one another and that equal weights are fair for each included individual model. Second, pooling all models with equal weights outperforms the best

individual model, and the result is robust for both portfolios and different estimation window schemes. This finding is consistent with the existing literature documenting that combining predictions from alternative models often improves upon those forecasts based on a single best model, particularly in an environment with individual models subject to structural breaks and misspecifications to varying degrees. Finally, we find that pooling models after optimal trimming always outperforms the pooling of all models for the 30 industry portfolios, and the pooling model with both optimal trimming and MCS trimming perform better than model pooling with the full set of models for the 25 size and B/M portfolios.

To consider model forecasting accuracy based on the different preferences of investors, Table 5, Table 6 and Table 7 report the results for different forecasting schemes with a particular focus on the central tendency, the left tail and the right tail. Consistent with the results for the entire portion of the density, model pooling after either MCS trimming or optimal trimming performs better than the best individual model and full model pooling, with several exceptions in the central tendency and left tail forecasts in which the MCS trimming performs worse than the full model pooling. The MCS trimming performs best in the central tendency forecast for both the 30 industry portfolios and the 25 size and B/M portfolios with the full sample, and optimal trimming performs best in all the other cases.

[Insert Table 4 here.]

[Insert Table 5 here.]

[Insert Table 6 here.]

[Insert Table 7 here.]

To better understand the forecasting scheme with model trimming and

pooling, for each case we present the selected models based on the best trimming scheme that maximizes the LPS of the density forecast, as shown in Figure 2. As the performance of asset pricing models is time-varying and as the true asset return-generating process may also change as a result of unobservable structural changes, we expect the selected models and their total numbers to change over time. All the plots in Figure 2 are generated based on rolling samples with a window size of 120 observations to capture the time-varying property. Panel A displays the results using 30 industry portfolios, and Panel B shows the results for 25 size and B/M portfolios. The figures show that when forming expectations for the central tendency and the right tail, the time variation of the number of models selected is much larger than the variation in the number of models for the left tail and the entire density. In the case of the 30 industry portfolios, although the number of selected models ranges from 2 to approximately 90 for the right tail and ranges from 2 to 50 for the central tendency, the number of models selected for the left tail and the entire density ranges only from approximately 30 to 75. A similar patten is observed in 25 size and B/M portfolios. The number of selected models for the entire density, the central tendency, the left tail and the right tail ranges from 37 to 27, 2 to 38, 37 to 27 and 14 to 38, respectively. Notably, for the central tendency and the right tail forecasts, the number of models selected increases significantly in the period surrounding recession periods, as defined by the National Bureau of Economic Research (NBER). In the figure, periods of economic stress are indicated by shaded bars, which implies that the underlying asset return-generating process becomes more complex and the model uncertainty more severe during times of economic stress and that more models must

therefore be selected into the model pool to address model uncertainty. By contrast, the number of selected models for the entire density and the left tail is relatively stable. These models show variation at the beginning of the sample and remain at a certain level throughout the remainder of the sample period.

[Insert Figure 2 here.]

To determine whether the numerical comparison is statistically significant, we employ the weighted likelihood ratio (WLR) test developed by Giacomini and White (2006) and Amisano and Giacomini (2007) to compare the forecast accuracy for each forecasting scheme. The null hypothesis is that the two competing models are equally accurate. We use an unweighted version of the WLR test to avoid imposing prior knowledge on the choice of a particular weighting function.

[Insert Table 8 here.]

Table 8 presents the results of these WLR tests, with Panels A, B, C, and D reporting results for the entire density, the central tendency, and the left tail and right tails, respectively. For the case with 30 industry portfolios as test assets, the optimal trimming clearly offers a statistically significant improvement relative to the best individual model, the best model set and the full model pool at more than a small significance level for the forecast of the entire density and the two tails. By contrast, the best model set and model pooling perform significantly better than the best individual model for the three types of forecasts. Whereas the best model set exhibits significant improvement relative to the model pool for the right tail forecast, the latter offers significant improvement over the former for the entire density and the left tail forecast. For the comparisons with respect to central ten-

dency forecasting, all the forecasting schemes perform quite similarly and show no significant difference. For the case with 25 size and B/M portfolios as test assets, the results are qualitatively similar to those in the 30 industry portfolios. Again, model trimming before pooling improves performance in a statistically significantly manner in all cases except for the central tendency forecast. Notably, the best model set offers further improvements over model pooling in these cases.

In summary, we show that trimming before model pooling is advantageous relative to pooling without trimming and model selection, particularly in an asset pricing context in which potential model specifications are rich. To uncover the underlying reason why the trimming scheme works best, we provide a statistical and economic explanation for the results in the following subsections. Further, we explore the economic implications of the trimming strategy by mimicking the decision-making process of an investor in actual practice.

6. Statistical and economic interpretations of the gains from trimming

We provide statistical explanations for the relatively good out-of-sample performance of model trimming. Using forecast encompassing tests, we demonstrate that trimming removes models that do not provide incremental information. We also show that trimming reduces estimation noise, thereby improving forecasting performance in terms of LPS metrics. In addition, we analyze the economic benefit of model trimming before combination by attempting to answer the question of whether the benefits vary over time and particularly across the business cycle and whether reliance on model

trimming leads to economically meaningful gains from the perspective of real investment practice.

6.1. Statistical explanation

As documented in Timmermann (2006), forecasts that add only marginal information should be excluded from the combination because the cost of their inclusion–increased parameter estimation error–is not matched by similar benefits. Before omitting these models, we need a statistical tool to determine the incremental information provided by each forecast and deciding which forecast can be removed. The forecast encompassing test provides a means for comparing the information content in different model forecasts. We employ the test statistics developed by Harvey, Leybourne, and Newbold (1998) and adapt them to the density forecast loss function, LPS, to make pairwise comparisons among the 120 asset pricing models. We let $d_{t+1} = (e_{i,t+1} - e_{j,t+1})e_{i,t+1}$, $\overline{d} = 1/(T-\tau)\sum_{t=\tau+1}^{T} d_t$, where $e_{i,t+1} = -LPS_{i,t+1}$ and $e_{j,t+1} = -LPS_{j,t+1}$. The test statistic is defined as

$$HLN = \frac{T-\tau-1}{T-\tau}\hat{V}(\overline{d})^{-1/2}\overline{d}, \tag{6.1}$$

where $\hat{V}(\overline{d}) = (T-\tau)^{-2}\sum_{t=\tau+1}^{T}(d_t - \overline{d})^2$. The test statistic follows a t-distribution with $T-\tau-1$ degrees of freedom, and we test the null hypothesis that the model A forecast encompasses the model B forecast against the (one-sided) alternative hypothesis that the model A forecast does not encompass the model B forecast. Therefore, if we reject the null hypothesis of encompassing, then it is useful to combine forecasts from models A and B rather than relying solely on the model A forecast.

Table 9 reports p-values for the encompassing test applied to the 120

asset pricing models in terms of out-of-sample entire density forecasting for the 30 industry portfolios and the 25 size and B/M portfolios based on the entire sample. Because of space limitations, we report only a portion of the representative results, and the full set of results is available upon request. Each number in the table corresponds to the null hypothesis that the forecast from the model in the row heading encompasses the forecast from the model in the column heading. Apparently, the encompass test cannot reject the null hypothesis in many cases. For example, consider the unconditional CAPM model in the first row of the table for the 30 industry portfolios; it encompasses the unconditional consumption CAPM model and the CRR5 model as well as all the conditional asset pricing models. The unconditional FF5 model listed in the sixth row of the table encompasses all models except for the unconditional Carhart4 model. These encompassing tests thus suggest that it is worthwhile to trim forecasts from individual models that do not provide additional information. This finding helps to explain the out-of-sample forecast gain from model trimming documented in Section.

[Insert Table 9 here.]

6.2. The sensitivity of the results to the trimming percentage

Although the encompassing test verifies the necessity of forecast trimming, another question arises regarding how many models we should trim to form the optimal forecast combination pool. The existence of pure noise forecasting in the model pool can inflate forecast error and thus generate a traditional type of bias-variance tradeoff. The highest possible number of models should be included to reduce bias. However, adding pure noise

or forecasts that only add marginal information will increase the variance in the forecast errors. A useful tool to assess this tradeoff is the sensitivity plot that depicts the LPS of density forecast as a function of the model trimming percentage.

Figure 3 shows how the LPS of the entire density forecast changes with the significance level of the MCS test and the trimming percentage in optimal trimming schemes using the entire sample. Panels A and B present the results for the 30 industry portfolios and 25 size and B/M portfolios, respectively. In the case using optimal trimming schemes to determine the final model combination pool, for the 30 industry portfolios, the LPS of the entire density forecast from the trimmed model pool increases when the trimming percentage changes from 0.1% to 50%, implying that trimming more forecasts from the bottom 50% of the models helps improve the forecast performance of the model pool. However, the LPS subsequently decreases when the trimming percentage varies from 50% to 99.9%, which thus indicates that trimming forecasts from the top 50% of the models causes the model pool performance to decline. The LPS achieves its highest level near -57,000, when almost all of the bottom 50% models are trimmed.

[Insert Figure 3 here.]

By contrast, the LPS constantly decreases with the increasing significance level in the MCS test. We know that choosing a low significance level in the MCS test will result in fewer models being trimmed, whereas a high significance level induces more models to be trimmed. This result thus suggests a soft trimming rule for the MCS scheme. Notably, the significance level α used in the MCS test does not imply that the bottom α% models are trimmed. For example, when $\alpha = 0.1\%$, we actually trim the

bottom 97% of the models. A similar patten is observed for the 25 size and B/M portfolios. The LPS of the density forecast reaches its highest level when trimming the bottom 20% of models (approximately) based on an optimal trimming scheme and when using a 0.1% significance level (trimming around the bottom 96% of the models) based on the MCS test. These sensitivity plots further support the benefit of model trimming in out-of-sample density forecasts.

6.3. Economic interpretation

We link the gains of model trimming in forming asset return density expectations to the real economy. First, we investigate whether the forecast gains from model trimming vary over time and are more prominent during the recession period. The intuition is that when the market is in distress, model uncertainty is more severe, which in turn leads to the failure of individual models and increased forecast noise, such that the benefits of model trimming are likely to be more pronounced in periods when the market is down. To explore these issues, we examine the performance of optimal trimming schemes relative to the best individual model (the Fama-French five-factor model) and full model pooling.

Figure 4 presents the difference in LPS of the entire density forecast between optimal trimming and best individual model/full model pooling based on a rolling sample. Panel A shows the results for the 30 industry portfolios, and Panel B provides the results for the 25 size and B/M portfolios. The top of each panel shows the difference between the optimal trimming and the best individual model, and the bottom of each panel shows the difference between the optimal trimming and the best model set

constructed based on the MCS test. In general, the difference in LPS between the optimal trimming and the best individual model (and full model pooling) is positive over time. In particular, the model trimming strikingly outperforms the other two schemes during periods of recession, as defined by the NBER (these periods are indicated by the shaded areas of the figure). These results suggest that model trimming is potentially an advantageous approach relative to model selection and model pooling in the context of asset pricing models, particularly during economic downturns.

[Insert Figure 4 here.]

6.4. Economic implications

Thus far, we have indicated the statistical significance of the trimming approach with respect to out-of-sample density expectations. However, examining whether model trimming leads to economically significant gains from the perspective of a real investor is of greater interest. We investigate the benefits of trimming using economic metrics to complement our statistical examinations. Density forecasting is more meaningful and relevant for asset allocation, and we evaluate the economic performance of trimming from the perspective of mean variance optimizing investors. We compare the realized certainty equivalent return (CER) and the Sharpe ratio using different asset return density forecasting methods: single best, model pooling with equal weights, model pooling with optimal weights, MCS trimming and optimal fixed trimming with both equal and optimal weights. We consider different specifications of investor risk aversions and transaction cost.

The decision problem of an investor can be generally described as follows:

$$\max_{\omega} \omega' \mathbf{Er}_t - (\frac{1}{2}\lambda)\omega' \mathbf{\Sigma}_t \omega - \sum_{i=1}^{m} \kappa(\omega_{i,t} - \omega_{i,t-1}) \tag{6.2}$$
$$s.t. \sum_{i=1}^{m} \omega_{i,t} \leq 1, \quad \omega_{i,t} \geq 0,$$

where ω denotes the $m \times 1$ vector of weights allocated to m risky assets by the investor at the beginning of month t, and $\mathbf{Er}_t$ and $\mathbf{\Sigma}_t$ are the vectors of expected excess returns and the covariance matrix of returns, respectively, which are extracted from the first 2 moments of the predictive return distribution obtained from a particular forecasting method documented in the preceding sections based on information available at the end of month $t-1$. In addition, λ represents the investor's degree of relative risk aversion, and κ is the proportional transaction cost incurred.

The CER of a portfolio can be calculated as $CER = E(r_p) - \frac{1}{2}\lambda\sigma_p^2$. The $E(r_p)$ is the mean expected return of the portfolio, which is calculated as $E(r_p) = \sum_{i=1}^{m} \omega_{i,t-1} Er_{i,t} + (1 - \sum_{i=1}^{m} \omega_{i,t}) r_{f_t}$, where r_f is the risk-free rate. σ_p is the variance in the portfolio obtained from $\sum_{i=1}^{m} \omega_{i,t-1}\sigma_{i,t} + \sum_{i=1}^{m} \omega_{i,t-1}(Er_i, t - E(r_p))(Er_i, t - E(r_p)'$. The Sharpe ratio can be obtained as follows: $Sharpe\ ratio = \frac{E(r_p)}{\sigma_p}$

Following O'Doherty, Savin, and Tiwari (2012), we also check the robustness of the results using 3 levels of investor risk aversion: (1) low risk aversion with λ equal to 2, (2) moderate risk aversion with λ equal to 5, and (3) high risk aversion with λ equal to 10. We also begin with a 0 transaction cost and then increase the value of κ to a more realistic number, 0.0025, which is equivalent to a proportional transaction cost of 50 basis points (bp) for a round-trip trade in stocks, and we assume that the investor's position

in the risk-free asset can be altered with no transaction costs.

We begin with the initial case without transaction costs, as presented in Table 10, which shows that the MCS and optimal trimming deliver the highest CERs and Sharpe ratio for investors with high risk aversion, with a λ equal to 10. For moderate risk aversion, the MCS and optimal trimming also have the highest economic gains under the full sample and rolling window forecasting scheme, whereas model pooling with equal weights using the expanding window performs better when investing in the 25 industry portfolios in addition to T-bills. For low risk aversion, the MCS and optimal trimming also generally perform better, except for the few cases involving investing in the 25 size and B/M portfolios. We then consider more realistic scenarios with transaction costs in Table 11. Considering the transaction costs, the single best model performs better than either pooling or trimming in only a few cases; this result is consistent with O'Doherty, Savin, and Tiwari (2012). Using the full sample as the forecasting window scheme and the strategy that invests in the 30 industry portfolios in addition to T-bills, we find that the MCS and optimal trimming obtain the best economic performance when evaluated using both the CER and Sharpe ratio under all levels of risk aversion.

[Insert Table 10 here.]

[Insert Table 11 here.]

7. Robustness Check

We now conduct a robustness check to ensure that the results are robust to the choices of the sample split points and different subsample periods. We perform robustness checks using 10 years and 25 years as the initial

estimation sample, in contrast to the main results based on five years. The 10-year and 25-year split points respectively contain 120 and 300 monthly observations in the model estimation period and leave the rest for out-of-sample evaluation.

These robustness check results are shown in Table 12. We can observe that the ranking of the four forecasting schemes remains using the two different split points. The optimal trimming scheme still performs best and is followed by the best model set constructed based on the MCS test, the model pool and the best individual model. The gains from model trimming are preserved even when the split point is changed.

[Insert Table 12 here.]

Additionally, in the previous analyses, we compare the forecasting performance of different schemes using the full out-of-sample periods. However, the four schemes might perform differently, and their ranking may differ over the various subsample evaluation periods. Thus, we again perform the rolling estimation using data from the first five years as the initial window and move one month ahead each time to re-estimate the model and to generate one-step-ahead forecasts. The out-of-sample forecasts are then compared in five different subsample periods: 1968-1978, 1978-1998, 1998-2008, and 2008-2011.

The results of the forecasting scheme comparisons in different evaluation periods in terms of the entire density forecast are presented in Table 13. The model weights in the forecast combination are generated based on the entire sample. The table format is similar to the previous tables, except that we compare five approximately 10-year evaluation periods. All the forecasting schemes perform at their worst in the 2008-2011 subperiod and at their

best in the 1968-1978 subperiod in the case using 30 industry portfolios as a testing asset, whereas in the case of 25 size and B/M portfolios, they perform the worst in the 1998-2008 subperiod and best in the 2008-2011 subperiod. More importantly, optimal trimming performs the best for the 30 industry portfolios in all subperiods, and the best model set constructed using MCS outperforms all others in the 25 size and B/M portfolios. Thus, model trimming generates forecasting gains not only for the entire sample but also for all the subperiods considered here.

[Insert Table 13 here.]

8. Conclusion

This paper compares out-of-sample density forecasts from 120 empirical asset pricing model comparisons. The paper not only considers a comprehensive set of models and evaluates density forecasts but also incorporates the accuracy preferences of different regions of the asset return distributions. The incorporation of distributional accuracy preferences distinguishes this paper from existing studies with asset pricing model comparisons, and this paper therefore has more practical value. The paper finds that although unconditional models are superior under most distributional accuracy preferences, certain conditional model specifications perform similarly and cannot be significantly differentiated from unconditional models when the central tendency or the right tail of the asset returns are weighted as more important. One potential explanation for the results is that the sample data that researchers in this field can normally obtain are not sufficiently informative to exclude conditional models from the model confidence set. Furthermore, conditional models can be useful in capturing particular parts of asset return

distributions. The findings in this regard can contribute new knowledge to the asset pricing model comparison literature.

The model comparison in this paper can provide useful guidance for practical use in selecting the best model. However, without observing the true asset pricing model, it is neither plausible to select a single model as the dominating model and to exclude potentially useful information contained in other models nor convincing to pool all models. In a model-rich context (i.e., when the number of models at one's disposal is large), it makes more sense to first trim models and then pool the sensible models. This paper shows that pooling the trimmed model set is advantageous regardless of which parts of the asset return distributions are regarded as more important. The statistical explanation of the results is that the trimmed models are the best encompassed by the selected models, and trimming exhibits more gains during economic distress when a fair number of models perform poorly. Notably, the advantages of trimming are shown to be both statistically significant and economically beneficial.

The usefulness of alternative asset pricing models has continued to pique the interest of scholars. As noted by Ferson, Nallareddy, and Xie (2013), the practical utility of an asset pricing model ultimately depends on its out-of-sample performance. The out-of-sample performance of density forecasting is of great interest, and most practical applications rely on this approach. Recent studies assessing alternative asset pricing models (e.g., Greenwood and Shleifer, 2014; Berk and van Binsbergen, 2015) include external information on investors' preferences to validate the models. Although these assessments aim to focus on validation beyond the practical utility of asset pricing models, it is also of practical importance to reflect investors' prefer-

ence in out-of-sample asset pricing model comparisons as well as to design ways to improve the model performance. Our study attempts to contribute to this line of research, and we expect further research in this direction in the future.

AppendixA. Description of asset pricing models

We include eight base asset pricing models. The data sources to construct each model are as follows:

(1) CAPM. The pricing factor of the CAPM is the market factor, and we use the value-weighted combined NYSE-AMEX-NASDAQ index provided by CRSP;

(2) CCAPM. The factor is the growth rate of real non-durables consumption, as reported by the Bureau of Economics Analysis, U.S. Department of Commerce;

(3) CRR5. Merton's 1973 intertemporal capital asset pricing model (ICAPM) notes that any state variable that predicts future investment opportunities serves as a state variable. Chen, Roll, and Ross (1986) use five macroeconomic variables (monthly growth in industrial production, change in expected inflation, unexpected inflation, risk premium, and a term structure factor) as pricing factors. Following O'Doherty, Savin, and Tiwari (2012), we collect the data to construct these five factors from the Ibbotson Yearbook and Laura Xiaolei Liu's website;

(4) JW. The JW model includes a CAPM market factor, the default premium, and the per capita labor income growth rate. We use the same market factor for the JW model as for the CAPM; we calculate the default premium as the lagged yield spread between BAA- and AAA-rated corporate bonds from the Board of Governors of the Federal Reserve System; we measure the per capita labor income growth rate

as $(L_{t-1} + L_{t-2})/(L_{t-2} + L_{t-3}) - 1$, where L is defined as the difference between personal income and dividend income per capita (from the Bureau of Economics Analysis, U.S. Department of Commerce). Following Jagannathan and Wang (1996), we calculate the 2-month moving average of the per capita labor income growth rate to reduce the influence of measurement errors;

(5) FF3. The FF3 augments the CAPM with with two additional factors: SMB, the return difference between small and large portfolios, and HML, the return difference between portfolios with high and low book-to-market ratios. We collect the relevant factor data from Kenneth French's website;

(6) Carhart4. The four-factor model introduces one more momentum factor, up minus down, to the Fama and French (1993) three-factor model, which can also be obtained from Kenneth French's website;

(7) FF5. The five-factor model adds two additional factors to the FF3, the return spread between a 30-year Treasury bond and the one-month T-bill and the return spread between long-term corporate and long-term government bonds. The series of these two factors are from Ibbotson Associates;

(8) LIQ. Recent studies note that liquidity factors play an important role in asset pricing, and Liu (2006) augments CAPM with a liquidity factor. We would like to thank this author for providing the liquidity factor data.

We summarize the common sets of instrumental variables used in existing asset pricing studies and use these variables to specify our conditional models. With the choice of different sets of instruments, we obtain a total of 120 empirical asset pricing models. We index the models from 1 to 120 shown in Table A.1. In the table,"U" denotes the unconditional specification, while "C" denotes the conditional specification. "ab" denotes the intercept and beta, and "f" denotes the factor risk premium. The Arabic number denotes which of the three instrument sets is used. The Lag IP, JAN and the vector (containing T-bill, DivYld, Term, Junk) are numbered sequentially from 1 to 3. For example, the CAPM(Ca1Ub) denotes the CAPM model with a time-varying intercept and a constant beta. The time-varying intercept is estimated using the first instrument, Lag IP.

Table A.1: Empirical specifications and number indices of the asset pricing models

Unconditional Models		**Conditional Models**							
	Index	**With time-varying factors**	**Index**	**With time-varying intercepts**	**Index**	**With time-varying beta (No intercept)**	**Index**	**With time-varying intercepts and beta**	**Index**
CAPM(Uab)	17	CCAPM(UabCf)	79	CAPM(Ca3Ub)	9	FF5(Cb3)	3	FF5(Cab3)	1
LIQ(Uab)	29	CAPM(UabCf)	80	CAPM(Ca2Ub)	14	FF3(Cb3)	8	Carhart4(Cab3)	2
Carhart4(Uab)	36	CCAPM(CbCf)	81	CAPM(Ca1Ub)	16	LIQ(Cb3)	26	FF3(Cab3)	4
FF3(Uab)	37	CAPM(CbCf)	82	LIQ(Ca3Ub)	22	FF3(Cb1)	40	Carhart4(Cb3)	5
CAPM(Ub)	45	CCAPM(UbCf)	84	Carhart4(Ca3Ub)	23	FF3(Cb2)	41	LIQ(Cab3)	6
FF3(Ub)	50	CAPM(UbCf)	85	FF3(Ca3Ub)	24	LIQ(Cb1)	44	CAPM(Cab3)	7
LIQ(Ub)	51	CCAPM(CabCf)	86	LIQ(Ca2Ub)	25	LIQ(Cb2)	46	Carhart4(Cab2)	10
Carhart4(Ub)	52	CAPM(CabCf)	87	LIQ(Ca1Ub)	27	FF5(Cb2)	48	CAPM(Cab2)	11
FF5(Uab)	55	LIQ(UabCf)	90	Carhart4(Ca2Ub)	28	FF5(Cb1)	53	JW(Cab3)	12
FF5(Ub)	60	LIQ(UbCf)	91	FF3(Ca2Ub)	30	CRR5(Cb3)	68	CAPM(Cab1)	13
JW(Uab)	63	LIQ(CbCf)	92	Carhart4(Ca1Ub)	34	CRR5(Cb1)	94	Carhart4(Cab1)	15
JW(Ub)	66	LIQ(CabCf)	93	FF3(Ca1Ub)	35	CRR5(Cb2)	95	LIQ(Cab1)	18
CCAPM(Uab)	77	JW(UabCf)	96	FF5(Ca3Ub)	47			FF3(Cab2)	19
CCAPM(Ub)	83	FF3(UabCf)	97	FF5(Ca2Ub)	49			FF3(Cab1)	20
CRR5(Uab)	109	FF3(UbCf)	98	FF5(Ca1Ub)	54			LIQ(Cab2)	21
CRR5(Ub)	114	JW(UbCf)	99	JW(Ca3Ub)	59			FF5(Cab2)	31
		JW(CbCf)	101	JW(Ca2Ub)	61			CAPM(Cb3)	32
		FF3(CbCf)	102	JW(Ca1Ub)	62			FF5(Cab1)	33
		JW(CabCf)	103	CCAPM(Ca3Ub)	70			Carhart4(Cb2)	38
		FF3(CabCf)	104	CCAPM(Ca2Ub)	74			Carhart4(Cb1)	39
		Carhart4(UabCf)	105	CCAPM(Ca1Ub)	75			CAPM(Cb1)	42
		Carhart4(UbCf)	106	CRR5(Ca3Ub)	100			CAPM(Cb2)	43
		Carhart4(CbCf)	110	CRR5(Ca2Ub)	107			JW(Cb3)	56
		Carhart4(CabCf)	111	CRR5(Ca1Ub)	108			JW(Cab2)	57
		CRR5(UabCf)	112					JW(Cab1)	58
		FF5(UabCf)	113					JW(Cb2)	64
		CRR5(UbCf)	115					JW(Cb1)	65
		FF5(UbCf)	116					CRR5(Cab3)	67
		FF5(CbCf)	117					CCAPM(Cab3)	69
		CRR5(CbCf)	118					CCAPM(Cb3)	71
		CRR5(CabCf)	119					CCAPM(Cab2)	72
		FF5(CabCf)	120					CCAPM(Cab1)	73
								CCAPM(Cb2)	76
								CCAPM(Cb1)	78
								CRR5(Cab1)	88
								CRR5(Cab2)	89

References

Amisano, G., Giacomini, R., 2007. Comparing density forecasts via weighted likelihood ratio tests. Journal of Business & Economic Statistics 25 (2), 177–190.

Armstrong, J. S., 1989. Combining forecasts: The end of the beginning or the beginning of the end? International Journal of Forecasting 5 (4), 585–588.

Berk, J. B., van Binsbergen, J. H., 2015. Assessing asset pricing models using revealed preference. Journal of Financial Economics, In press.

Bollerslev, T., Todorov, V., Xu, L., 2015. Tail risk premia and return predictability. Journal of Financial Economics 118 (1), 113–134.

Box, G. E., 1980. Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society. Series A (General) 143, 383–430.

Breeden, D. T., 1979. An intertemporal asset pricing model with stochastic consumption and investment opportunities. Journal of Financial Economics 7 (3), 265–296.

Carhart, M. M., 1997. On persistence in mutual fund performance. The Journal of Finance 52 (1), 57–82.

Chang, B. Y., Christoffersen, P., Jacobs, K., 2013. Market skewness risk and the cross section of stock returns. Journal of Financial Economics 107 (1), 46–68.

Chen, N.-F., Roll, R., Ross, S. A., 1986. Economic forces and the stock market. Journal of Business, 383–403.

Diebold, F. X., Gunther, T. A., Tay, A. S., 1998. Evaluating density forecasts with applications to financial risk management. International Economic Review 39 (4), 863–883.

Diebold, F. X., Lopez, J. A., 1996. Forecast evaluation and combination. Working paper 192, National Bureau of Economic Research.
URL `http://www.nber.org/papers/t0192`

Fama, E., French, K., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33 (1), 3–56.

Fama, E. F., French, K. R., 1996. Multifactor explanations of asset pricing anomalies. The Journal of Finance 51 (1), 55–84.

Fama, E. F., French, K. R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Ferson, W., Nallareddy, S., Xie, B., 2013. The "out-of-sample" performance of long run risk models. Journal of Financial Economics 107 (3), 537–556.

Ghysels, E., 1998. On stable factor structures in the pricing of risk: Do time-varying betas help or hurt? The Journal of Finance 53 (2), 549–573.

Giacomini, R., White, H., 2006. Tests of conditional predictive ability. Econometrica 74 (6), 1545–1578.

González-Rivera, G., Lee, T. H., Mishra, S., 2004. Forecasting volatility: A reality check based on option pricing, utility function, value-at-risk, and predictive likelihood. International Journal of Forecasting 20 (4), 629–645.

Greenwood, R., Shleifer, A., 2014. Expectations of returns and expected returns. Review of Financial Studies 27 (3), 714–746.

Hansen, P. R., Lunde, A., Nason, J. M., 2003. Choosing the best volatility models: The model confidence set approach. Oxford Bulletin of Economics and Statistics 65 (1), 839–861.

Hansen, P. R., Lunde, A., Nason, J. M., 2011. The model confidence set. Econometrica 79 (2), 453–497.

Hansen, P. R., Timmermann, A., 2012. Choice of sample split in out-of-sample forecast evaluation. Working paper, ECO.
URL http://cadmus.eui.eu/handle/1814/21454

Harvey, D. I., Leybourne, S. J., Newbold, P., 1998. Tests for forecast encompassing. Journal of Business and Economic Statistics 16, 254–259.

Hodrick, R., Zhang, X., 2001. Evaluating the specification errors of asset pricing models. Journal of Financial Economics 62 (2), 327–376.

Jagannathan, R., Wang, Z., 1996. The conditional CAPM and the cross-section of expected returns. The Journal of Finance 51 (1), 3–53.

Kadan, O., Liu, F., 2014. Performance evaluation with high moments and disaster risk. Journal of Financial Economics 113 (1), 131–155.

Kan, R., Robotti, C., 2009. Model comparison using the Hansen-Jagannathan distance. Review of Financial Studies 22 (9), 3449–3490.

Kan, R., Robotti, C., Shanken, J., 2013. Pricing model performance and the two-pass cross-sectional regression methodology. The Journal of Finance 68 (6), 2617–2649.

Kelly, B., Jiang, H., 2014. Tail risk and asset prices. Review of Financial Studies 27 (10), 2841–2871.

Lewellen, J., Nagel, S., Shanken, J., 2010. A skeptical appraisal of asset pricing tests. Journal of Financial Economics 96 (2), 175–194.

Lintner, J., 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. The Review of Economics and Statistics 47, 13–37.

Liu, W., 2006. A liquidity-augmented capital asset pricing model. Journal of Financial Economics 82 (3), 631–671.

Merton, R. C., 1973. An intertemporal capital asset pricing model. Econometrica: Journal of the Econometric Society 41, 867–887.

Mossin, J., 1966. Equilibrium in a capital asset market. Econometrica: Journal of the Econometric Society 34, 768–783.

O'Doherty, M., Savin, N. E., Tiwari, A., 2012. Modeling the cross section of stock returns: A model pooling approach. Journal of Financial and Quantitative Analysis 47 (6), 1331–1360.

Rapach, D. E., Strauss, J. K., Zhou, G., 2010. Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. Review of Financial Studies 23 (2), 821–862.

Sharpe, W. F., 1964. Capital asset prices: A theory of market equilibrium under conditions of risk*. The Journal of Finance 19 (3), 425–442.

Simin, T., 2008. The poor predictive performance of asset pricing models. The Journal of Financial and Quantitative Analysis 43 (2), pp. 355–380.

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

Timmermann, A., 2006. Forecast combinations. Handbook of Economic Forecasting 1, 135–196.

Table 1: Summary statistics for log predictive scores for individual asset pricing models

This table reports the descriptive statistics for the (weighted) log predictive scores (denoted as LPSs or WLPSs) for the 120 asset pricing models. The models included are listed in AppendixA. For a given model, the WLPS is computed from the predictive densities based on equation (3.11) using the full sample. The table reports results for two distinct sets of test assets, namely, the 30 value-weighted industry portfolios and the standard Fama-French 25 size and B/M portfolios. The four panels represent four different distributional accuracy preferences. Panels A to D present results for equal weighting, the central tendency accuracy preference, and the left and right tail preferences, respectively. Each panel summarizes the maximum, minimum, 75%, 50%, and 25% quartiles and mean of the (W)LPS of the all the models included.

	30 industry portfolios		25 size and B/M portfolios	
Panel A: Equal				
	LPS	Model	LPS	Model
Max	-58806.5269	FF5(Ub)	-28337.6825	FF3(Ub)
Min	-88888.5509	CRR5(Cab3)	-67155.8090	CRR5(Cab3)
75\%	-68438.2414	CCAPM(Ca1Ub)	-51279.7880	CAPM(UabCf)
50\%	-66657.6769	CRR5(Ub)	-42337.2928	JW(Cab1)
25\%	-61657.6868	FF3(Ca3Ub)	-37657.5897	FF5(Cb3)
Mean	-65685.7382		-43194.5099	
Panel B: Central tendency				
	WLPS	Model	WLPS	Model
Max	-5.9199E-25	JW(Ub)	-1.0103E-12	LIQ(Cb3)
Min	-1.1296E-24	CRR5(Cab3)	-2.9887E-12	CRR5(CabCf)
75\%	-7.1124E-25	CCAPM(Ub)	-2.0555E-12	CRR5(Ub)
50\%	-6.9636E-25	LIQ(UabCf)	-1.5957E-12	JW(Uab)
25\%	-6.3344E-25	FF3(Ca1Ub)	-1.1784E-12	FF5(Ca1Ub)
Mean	-6.8666E-25		-1.6465E-12	
Panel C: Left tail				
	WLPS	Model	WLPS	Model
Max	-55728.2537	FF5(Ub)	-22514.7137	FF3(Ub)
Min	-84263.4497	CRR5(Cab3)	-53138.0808	CRR5(Cab3)
75\%	-64287.9939	CRR5(Ca2Ub)	-40606.1465	CAPM(UabCf)
50\%	-63091.0752	CRR5(Ub)	-33628.0721	JW(Cab1)
25\%	-58466.2133	FF3(Ca3Ub)	-29811.2212	FF5(Cb3)
Mean	-62226.0072		-34241.2057	
Panel D: Right tail				
	WLPS	Model	WLPS	Model
Max	-3063.0472	FF3(Ub)	-5822.9689	FF3(Ub)
Min	-4625.1012	CRR5(Cab3)	-14017.7282	CRR5(Cab3)
75\%	-3643.4304	CRR5(Ca2Ub)	-10642.9444	CRR5(Uab)
50\%	-3551.8357	FF5(UbCf)	-8709.2207	JW(Cab1)
25\%	-3190.7245	CAPM(Ca1Ub)	-7846.3684	FF5(Cb3)
Mean	-3459.7310		-8953.3042	

Table 2: Comparison of individual asset pricing models: out-of-sample point forecasts (1968-2011)

This table compares the out-of-sample mean return point forecast constructed from a proportion of the 120 asset pricing models described in AppendixA. Because of space limitations, the complete results for all models are not presented here but are available upon request from the authors. The models reported are representative of the main findings. The testing sample period is from 1968 to 2011. The table reports results for two distinct sets of test assets, namely, the 30 value-weighted industry portfolios and the standard Fama-French 25 size and B/M portfolios. Ave $|PE|$ is the average absolute pricing error, and Ave PE^2 represents the average squared pricing error. p_{mcs} denotes the p-value of the Hansen, Lunde, and Nason (2011) model confidence set (MCS) test at a significance level of 0.1%. The numbers in bold are the smallest number or the lowest loss in each column.

	Panel A: 30 industry portfolios				**Panel B: 25 size and B/M portfolios**			
Model	Ave $\|PE\|$	P_{mcs}	Ave PE^2	P_{mcs}	Ave $\|PE\|$	P_{mcs}	Ave PE^2	P_{mcs}
CAPM(U)	0.1620	0.0000	0.0501	0.0000	0.2648	0.0000	0.0963	0.0000
CCAPM(U)	0.1358	0.0000	0.0332	0.0000	0.1495	0.0000	0.0288	0.0000
CRR5(U)	0.2641	0.0000	0.0951	0.0000	0.2999	0.0000	0.1043	0.0000
Carhart4(U)	**0.0390**	**1.0000**	**0.0024**	**1.0000**	**0.0301**	**1.0000**	**0.0021**	**1.0000**
FF3(U)	0.4747	0.0000	0.2594	0.0000	0.5865	0.0000	0.3767	0.0000
FF5(U)	0.0515	0.0000	0.0037	0.4450	0.0777	0.0000	0.0085	0.0000
JW(U)	0.1675	0.0000	0.0466	0.0000	0.1018	0.0000	0.0191	0.0000
LIQ(U)	0.0472	0.7210	0.0032	0.5010	0.0781	0.0000	0.0093	0.0000
CAPM(C)	0.1835	0.0000	0.0566	0.0000	0.0978	0.0000	0.0207	0.0000
CCAPM(C)	0.0513	0.3330	0.0042	0.0020	0.0703	0.5020	0.0084	0.8270
CRR5(C)	0.2198	0.0000	0.0805	0.0000	0.0704	0.0000	0.0094	0.0000
Carhart4(C)	0.0980	0.0000	0.0138	0.0000	0.1504	0.0000	0.0278	0.0000
FF3(C)	0.0528	0.0000	0.0051	0.0000	0.0447	0.6110	0.0034	0.8270
FF5(C)	0.0632	0.0000	0.0055	0.0000	0.0773	0.0000	0.0086	0.0000
JW(C)	0.1674	0.0000	0.0534	0.0000	0.1267	0.0000	0.0240	0.0000
LIQ(C)	0.1648	0.0000	0.0332	0.0000	0.1161	0.0000	0.0165	0.0000

Table 3: Comparison of individual asset pricing models: out-of-sample tail interval forecasts

This table compares the out-of-sample tail interval forecast, the value-at-risk (VaR), constructed from a proportion of the 120 asset pricing models described in AppendixA. Because of space limitations, the complete results for all models are not presented here but are available upon request from the authors. The models reported represent the main findings. The testing sample period is from 1968 to 2011.The table reports results for two distinct sets of test assets, namely, the 30 value-weighted industry portfolios and the standard Fama-French 25 size and B/M portfolios. The VaR loss is defined and calculated from equation (3.14). p_{mcs} denotes the p-value of the Hansen, Lunde, and Nason (2011) model confidence set (MCS) test at a significance level of 0.1%. The numbers in bold are the smallest number or the lowest loss in each column.

	Panel A: 30 industry portfolios		**Panel B: 25 size and B/M portfolios**	
Model	VaR loss	P_{mcs}	VaR loss	P_{mcs}
CAPM(U)	0.9548	0.4900	1.0486	0.0390
CCAPM(U)	1.1714	0.0510	1.2554	0.0080
CRR5(U)	1.3935	0.0000	1.4926	0.0000
Carhart4(U)	1.2184	0.0510	1.2728	0.0080
FF3(U)	1.0416	0.0510	1.1163	0.0390
FF5(U)	1.0730	0.0510	1.1261	0.0080
JW(U)	**0.9433**	**1.0000**	**1.0256**	**1.0000**
LIQ(U)	1.1013	0.0510	1.1494	0.0080
CAPM(C)	0.9563	0.2920	1.0366	0.0390
CCAPM(C)	1.1143	0.0510	1.1333	0.0080
CRR5(C)	0.9788	0.0510	1.0356	0.0960
Carhart4(C)	1.0769	0.0510	1.1246	0.0390
FF3(C)	0.9857	0.0510	1.1105	0.0390
FF5(C)	1.1002	0.0510	1.1274	0.0080
JW(C)	0.9704	0.0510	1.0736	0.0390
LIQ(C)	1.1575	0.0510	1.2173	0.0080

Table 4: Comparison of model trimming and pooling without accuracy preferences

This table compares out-of-sample density forecasts constructed from the single best model and three types of model pools, namely, the model pool with a full set of models (Pool), the model pool with a trimmed model set using MCS trimming (MCS Best Model Set) and the optimal fixed trimming (Optimal Trimmed). The models are pooled using both simple equal weighting (EW) and optimal weighting (OptW). The table assumes that the investor weights different regions of the asset distribution equally, i.e., with no distributional preference. Panel A reports the log predictive scores (LPSs) for the 30 value-weighted industry portfolios, and Panel B presents LPSs for the standard Fama-French 25 size and B/M portfolios. The forecasts are made using the full sample, rolling window and expanding window. The numbers in bold are the highest LPSs in the column.

Method	Full sample	Rolling sample	Expanding sample
Panel A: 30 industry portfolios			
Best Individual Model	-58806.5269	-13576.5301	-35038.8957
Pool (EW)	-57249.1131	-13223.1479	-34134.9119
Pool (OptW)	-59715.0635	-13791.3497	-35650.3174
MCS Best Model Set (EW)	-57740.1320	-13276.9623	-34448.5022
MCS Best Model Set (OptW)	-57739.4894	-13276.8614	-34448.1943
Optimal Trimmed (EW)	-57040.7461	-13175.5444	-34004.8995
Optimal Trimmed (OptW)	**-57035.1348**	**-13174.4918**	**-34001.5961**
Panel B: 25 size and B/M Portfolios			
Best Individual Model	-28337.6825	-6455.2860	-16970.9397
Pool (EW)	-28011.2539	-6396.9988	-16808.9017
Pool (OptW)	-30343.7427	-6933.7763	-18241.1301
MCS Best Model Set (EW)	-27487.0239	-6273.6538	-16511.3390
MCS Best Model Set (OptW)	-27487.3096	-6273.6646	-16511.5821
Optimal Trimmed (EW)	-27342.0575	-6242.2426	-16413.5114
Optimal Trimmed (OptW)	**-27340.7510**	**-6241.1723**	**-16411.5909**

Table 5: Comparison of model trimming and pooling with accuracy preferences for the central tendency

This table compares out-of-sample density forecasts from the single best model and three types of model pools, namely, the model pool with full set of models (Pool), the model pool with trimmed model set using MCS trimming (MCS Best Model Set) and the optimal fixed trimming (Optimal Trimmed). The models are pooled using both simple equal weighting (EW) and optimal weighting (OptW). The table assumes that the investor weights the forecast accuracy of the central tendency of the asset distribution as more important. Panel A reports the weighted log predictive scores (WLPSs) for the 30 value-weighted industry portfolios, and Panel B presents WLPSs for the standard Fama-French 25 size and B/M portfolios. The forecasts are made using the full sample, rolling window and expanding window. The numbers in bold are the highest WLPSs in the column.

Method	Full sample	Rolling Sample	Expanding sample
Panel A: 30 industry portfolios			
Single Best Model	-5.9199E-25	-4.92E-25	-3.9576E-25
Pool (EW)	-5.9929E-25	-4.6374E-25	-3.017E-25
Pool (OptW)	-6.3221E-25	-4.8835E-25	-3.1858E-25
MCS Best Model Set (EW)	-5.8589E-25	-4.6374E-25	-3.017E-25
MCS Best Model Set (OptW)	**-5.8587E-25**	-4.6306E-25	-3.0132E-25
Optimal Trimmed (EW)	-5.884E-25	-4.4738E-25	-2.9679E-25
Optimal Trimmed (OptW)	-5.884E-25	**-4.4736E-25**	**-2.9678E-25**
Panel B: 25 size and B/M portfolios			
Single Best Model	-1.0103E-12	-2.8708E-13	-7.6878E-13
Pool (EW)	-1.078E-12	-9.3859E-14	-5.6756E-13
Pool (OptW)	-1.2349E-12	-1.0635E-13	-6.458E-13
MCS Best Model Set (EW)	-9.8801E-13	-9.3859E-14	-5.6756E-13
MCS Best Model Set (OptW)	**-9.8789E-13**	-9.2819E-14	-5.6142E-13
Optimal Trimmed (EW)	-1.0085E-12	**-8.7716E-14**	-5.3631E-13
Optimal Trimmed (OptW)	-1.0081E-12	-8.7717E-14	**-5.362E-13**

Table 6: Comparison of model trimming and pooling with accuracy preferences for the left tail

This table compares out-of-sample density forecasts from the single best model and three types of model pools, namely, the model pool with full set of models (Pool), the model pool with trimmed model set using MCS trimming (MCS Best Model Set) and the optimal fixed trimming (Optimal Trimmed). The models are pooled using both simple equal weighting (EW) and optimal weighting (OptW). The table assumes that the investor weights the forecast accuracy of the left tail of the asset distribution as more important. Panel A reports the weighted log predictive scores (WLPSs) for the 30 value-weighted industry portfolios, and Panel B presents the WLPSs for the standard Fama-French 25 size and B/M portfolios. The forecasts are made using the full sample, rolling window and expanding window. The numbers in bold are the highest WLPSs in the column.

Method	Full sample	Rolling sample	Expanding sample
Panel A: 30 industry portfolios			
Single Best Model	-55728.2537	-12970.7945	-33190.2797
Pool (EW)	-54281.9429	-12640.1063	-32341.7020
Pool (OptW)	-56618.2457	-13182.6582	-33776.3960
MCS Best Model Set (EW)	-54762.9739	-12696.1849	-32620.8011
MCS Best Model Set (OptW)	-54762.8647	-12696.0915	-32620.5558
Optimal Trimmed (EW)	-54086.6217	-12594.7428	-32220.1794
Optimal Trimmed (OptW)	**-54081.1989**	**-12593.7235**	**-32216.9643**
Panel B: 25 size and B/M portfolios			
Single Best Model	-22514.7137	-5156.2693	-13444.8753
Pool (EW)	-22250.3153	-5108.7521	-13310.9732
Pool (OptW)	-24102.3409	-5535.9947	-14450.4524
MCS Best Model Set (EW)	-21838.0966	-5006.8436	-13064.3019
MCS Best Model Set (OptW)	-21838.3545	-5006.8526	-13064.5108
Optimal Trimmed (EW)	-21715.8897	-4986.5199	-12991.5981
Optimal Trimmed (OptW)	**-21715.0386**	**-4985.6163**	**-12990.7743**

Table 7: Comparison of model trimming and pooling with accuracy preferences for the right tail

This table compares out-of-sample density forecasts from the single best model and three types of model pools, namely, the model pool with full set of models (Pool), the model pool with trimmed model set using MCS trimming (MCS Best Model Set) and the optimal fixed trimming (Optimal Trimmed). The models are pooled using both simple equal weighting (EW) and optimal weighting (OptW). The table assumes that the investor weights the forecast accuracy of the right tail of the asset distribution as more important. Panel A reports the weighted log predictive scores (WLPSs) for the 30 value-weighted industry portfolios, and Panel B presents WLPSs for the standard Fama-French 25 size and B/M portfolios. The forecasts are made using the full sample, rolling window and expanding window. The numbers in bold are the highest WLPSs in the column.

Method	Full sample	Rolling sample	Expanding sample
Panel A: 30 industry portfolios			
Single Best Model	-3063.0472	-595.2215	-1847.5006
Pool (EW)	-2967.1701	-583.0416	-1793.2099
Pool (OptW)	-3096.7025	-608.6125	-1873.8404
MCS Best Model Set (EW)	-2955.3241	-582.8607	-1792.2300
MCS Best Model Set (OptW)	-2954.7743	-582.5064	-1791.0964
Optimal Trimmed (EW)	-2953.1927	-579.5898	-1784.0969
Optimal Trimmed (OptW)	**-2952.9598**	**-579.5456**	**-1783.9720**
Panel B: 25 size and B/M portfolios			
Single Best Model	-5822.9689	-1287.1004	-3522.1637
Pool (EW)	-5760.9386	-1288.2467	-3497.9285
Pool (OptW)	-6241.4142	-1397.6818	-3790.5856
MCS Best Model Set (EW)	-5646.6626	-1259.0836	-3426.1081
MCS Best Model Set (OptW)	-5646.6391	-1257.8978	-3425.7746
Optimal Trimmed (EW)	-5626.1678	-1254.9516	-3420.3990
Optimal Trimmed (OptW)	**-5625.7244**	**-1254.8462**	**-3419.2161**

Table 8: Comparison of model trimming and pooling: weighted likelihood ratio tests

This table reports the results from the weighted likelihood ratio (WLR) tests developed by Giacomini and White (2006) and Amisano and Giacomini (2007). This test is used to compare the significance of the difference between two competing density forecasts. The one-sided test examines whether the predictive performance of forecast scheme A is significantly better than that of forecast scheme B. The out-of-sample density forecasts are constructed from the single best model and three types of model pools, namely, the model pool with full set of models (model pooling), the model pool with trimmed model set using MCS trimming (MCS Best Model Set) and the optimal fixed trimming (Optimal trimming). We use an unweighted version of the WLR test to avoid imposing prior knowledge on the choice of a particular weighting function. The table reports results for two distinct sets of test assets, namely, the 30 value-weighted industry portfolios and the standard Fama-French 25 size and B/M portfolios. Panel A assumes no distributional accuracy preferences, and Panels B to D assume that the investor respectively weights the central part, the left tail and the right tail of the asset return distribution as more important.

$H_A : E(\overline{WLR}_{A,B}) > 0$

Panel A: Equal

30 industry portfolios

Scheme A/Scheme B	Best Individual Model	MCS Best Model Set (best)	Model Pooling (best)
Optimal trimming (best)	12.4362(0.0000)	8.3017(0.0000)	7.6399(0.0000)
Model Pooling (best)	10.6264(0.0000)	4.9946(0.0000)	
MCS Best Model Set (best)	9.9717(0.0000)		

25 size and B/M portfolios

Scheme A/Scheme B	Best Individual Model	Model Pooling (best)	MCS Best Model Set (best)
Optimal trimming (best)	12.8618(0.0000)	30.9619(0.0000)	5.1933(0.0000)
MCS Best Model Set (best)	12.7717(0.0000)	15.1015(0.0000)	
Model Pooling (best)	4.0709(0.0000)		

Panel B: Central Tendency

30 industry portfolios

Scheme A/Scheme B	Best Individual Model	MCS Best Model Set (best)	Model Pooling (best)
MCS Best Model Set (best)	1.4845(0.0691)	1.2470(0.1065)	0.6734(0.2505)
Optimal trimming (best)	0.8795(0.1898)	0.8250(0.2049)	
Model Pooling (best)	0.7333(0.2318)		

25 size and B/M portfolios

Scheme A/Scheme B	Best Individual Model	Model Pooling (best)	Optimal trimming (best)
MCS Best Model Set (best)	0.8650(0.1937)	1.0276(0.1523)	0.9370(0.1746)
Optimal trimming (best)	0.0413(0.4835)	1.0573(0.1454)	
Model Pooling (best)	0.6359(0.2625)		

Table 8 continued on next page

Table 8 continued

Panel C: Left Tail			
30 industry portfolios			
Scheme A/Scheme B	Best Individual Model	MCS Best Model Set (best)	Model Pooling (best)
Optimal trimming (best)	12.1831(0.0000)	7.9329(0.0000)	7.5265(0.0000)
MCS Model Pooling (best)	10.0718(0.0000)	5.0433(0.0000)	
Best Model Set (best)	10.4383(0.0000)		
25 Size and B/M Portfolios			
Scheme A/Scheme B	Best Individual Model	Model Pooling (best)	MCS Best Model Set (best)
Optimal trimming (best)	12.8618(0.0000)	30.9619(0.0000)	5.1933(0.0000)
MCS Best Model Set (best)	12.7717(0.0000)	15.1015(0.0000)	
Model Pooling (best)	4.0709(0.0000)		

Panel D: Right Tail			
30 industry portfolios			
Scheme A/Scheme B	Best Individual Model	MCS Best Model Set (best)	Model Pooling (best)
Optimal trimming (best)	12.1831(0.0000)	7.9329(0.0000)	7.5265(0.0000)
MCS Best Model Set (best)	10.0718(0.0000)	5.0433(0.0000)	
Model Pooling (best)	10.4383(0.0000)		
25 size and B/M portfolios			
Scheme A/Scheme B	Best Individual Model	Model Pooling (best)	MCS Best Model Set (best)
Optimal trimming (best)	11.9349(0.0000)	25.2746(0.0000)	5.0371(0.0000)
MCS Best Model Set (best)	11.7400(0.0000)	13.2077(0.0000)	
Model Pooling (best)	3.8961(0.0000)		

Table 9: Forecast encompassing test

This table reports p-values for the Harvey, Leybourne, and Newbold (1998) MHLN statistic. The models included are listed in AppendixA. Panel A presents results using the 30 value-weighted industry portfolios, while Panel B reports the results for the standard Fama-French 25 size and B/M portfolios. The null hypothesis is that the forecast given in the row heading encompasses the forecast given in the column heading against the alternative hypothesis that the forecast given in the column heading does not encompass the forecast given in the row heading.

Panel A: 30 industry portfolios

	CAPM(U)	CCAPM(U)	CRR5(U)	Carhart4(U)	FF3(U)	FF5(U)	JW(U)	LIQ(U)	CAPM(C)	CCAPM(C)	CRR5(C)	Carhart4(C)	FF3(C)	FF5(C)	JW(C)	LIQ(C)
CAPM(U)		0.0000	0.0000	1.0000	1.0000	1.0000	0.7280	0.9971	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CCAPM(U)	1.0000		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9920	0.7303	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
CRR5(U)	1.0000	0.0000		1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.7344	0.2438	0.0800	0.6351	0.0806	0.0041
Carhart4(U)	0.0000	0.0000	0.0000		0.0386	0.9492	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
FF3(U)	0.0000	0.0000	0.0000	0.9614		0.9998	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
FF5(U)	0.0000	0.0000	0.0000	0.0508	0.0002		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
JW(U)	0.2720	0.0000	0.0000	1.0000	1.0000	1.0000		0.8466	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
LIQ(U)	0.0029	0.0000	0.0000	1.0000	1.0000	1.0000	0.1534		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CAPM(C)	1.0000	0.0080	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000		0.0055	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
CCAPM(C)	1.0000	0.2697	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9945		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
CRR5(C)	1.0000	0.0000	0.2656	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000		0.0005	0.0001	0.2713	0.0007	0.0000
Carhart4(C)	1.0000	0.0000	0.7562	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.9995		0.0541	0.9880	0.1544	0.0020
FF3(C)	1.0000	0.0000	0.9200	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.9999	0.9459		0.9985	0.5554	0.0085
FF5(C)	1.0000	0.0000	0.3649	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.7287	0.0120	0.0015		0.0073	0.0000
JW(C)	1.0000	0.0000	0.9194	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.9993	0.8456	0.4446	0.9927		0.0126
LIQ(C)	1.0000	0.0000	0.9959	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	1.0000	0.9980	0.9915	1.0000	0.9874	

Panel B: 25 size and B/M portfolios

	CAPM(U)	CCAPM(U)	CRR5(U)	Carhart4(U)	FF3(U)	FF5(U)	JW(U)	LIQ(U)	CAPM(C)	CCAPM(C)	CRR5(C)	Carhart4(C)	FF3(C)	FF5(C)	JW(C)	LIQ(C)
CAPM(U)		0.0000	0.0000	1.0000	1.0000	1.0000	0.2858	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CCAPM(U)	1.0000		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9956	0.4725	0.9998	1.0000	0.9999	1.0000	0.9619	0.9990
CRR5(U)	1.0000	0.0000		1.0000	1.0000	1.0000	1.0000	1.0000	0.0002	0.0000	0.1081	0.1661	0.0269	0.2663	0.0004	0.0020
Carhart4(U)	0.0000	0.0000	0.0000		0.9820	0.2891	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
FF3(U)	0.0000	0.0000	0.0000	0.0180		0.0534	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
FF5(U)	0.0000	0.0000	0.0000	0.7109	0.9466		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
JW(U)	0.7142	0.0000	0.0000	1.0000	1.0000	1.0000		1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
LIQ(U)	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000	0.0000		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CAPM(C)	1.0000	0.0044	0.9998	1.0000	1.0000	1.0000	1.0000	1.0000		0.0004	0.9943	0.9999	0.9975	0.9998	0.4905	0.9469
CCAPM(C)	1.0000	0.5275	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9996		1.0000	1.0000	1.0000	1.0000	0.9891	1.0000
CRR5(C)	1.0000	0.0002	0.8919	1.0000	1.0000	1.0000	1.0000	1.0000	0.0057	0.0000		0.7984	0.1947	0.9117	0.0004	0.0241
Carhart4(C)	1.0000	0.0000	0.8339	1.0000	1.0000	1.0000	1.0000	1.0000	0.0001	0.0000	0.2016		0.0112	0.7329	0.0000	0.0012
FF3(C)	1.0000	0.0001	0.9731	1.0000	1.0000	1.0000	1.0000	1.0000	0.0025	0.0000	0.8053	0.9888		0.9775	0.0002	0.0389
FF5(C)	1.0000	0.0000	0.7337	1.0000	1.0000	1.0000	1.0000	1.0000	0.0002	0.0000	0.0883	0.2671	0.0225		0.0000	0.0018
JW(C)	1.0000	0.0381	0.9996	1.0000	1.0000	1.0000	1.0000	1.0000	0.5095	0.0109	0.9996	1.0000	0.9998	1.0000		0.8931
LIQ(C)	1.0000	0.0010	0.9980	1.0000	1.0000	1.0000	1.0000	1.0000	0.0531	0.0000	0.9759	0.9988	0.9611	0.9982	0.1069	

Table 10: Economic implication: asset allocation without transaction cost

The table compares the out-of-sample performance of alternative methods in addressing asset pricing model uncertainty in terms of monthly certainty equivalent rates of return (CERs) expressed as a percentage and Sharpe ratio (Sharpe) for mean-variance optimal investment strategies. The asset allocation practice uses the moments of stock return predictive distributions implied by each of the forecasting schemes to choose an optimal optimal portfolio allocation from 1978 to 2011 using three different window schemes: the full sample, rolling window and expanding window. The asset universe includes of T-bills and i) the 30 value-weighted industry portfolios or ii) the 25 size and book-to-market (B/M) portfolios. The results are presented for three different levels of risk aversion: 2, 5 and 10. The forecasting scheme includes the single best model and three types of model pools, namely, the model pool with full set of models (Pool), the model pool with trimmed model set using MCS trimming (MCS Best Model Set) and the optimal fixed trimming (Optimal Trimmed). The EW and OptW in brackets represent pooling using equal and optimal weights, respectively. This table assumes no transaction costs. The number in bold indicates the highest economic gain in the column.

Risk aversion	**Forecasting schemes**	**Full sample**				**Rolling window**				**Expanding window**			
		30		25		30		25		30		25	
		CER	**Sharp**	**CER**	**Sharp**	**CER**	**Sharp**	**CER**	**Sharp**	**CER**	**Sharp**	**CER**	**Sharp**
2	**Single best model**	1.7094	0.4592	2.0119	0.7461	0.9402	0.1809	0.8978	0.1744	0.9587	0.1850	0.9141	0.1806
2	**Pool (EW)**	2.0064	0.4385	**2.4402**	0.7003	0.9939	0.1838	**0.9942**	0.2046	**0.9939**	0.1838	**0.9942**	**0.2046**
2	**Pool (OptW)**	0.6894	0.1171	0.6720	0.1771	0.6894	0.1171	0.6720	0.1771	0.6894	0.1171	0.6720	0.1771
2	**MCS Best model set (EW)**	1.8493	0.6519	1.9872	0.7789	0.9011	0.1827	0.8675	0.1791	0.9134	0.1863	0.8909	0.1923
2	**MCS Best model set (OptW)**	1.8507	**0.6772**	1.9773	**0.7945**	0.8995	0.1809	0.8862	0.1811	0.9405	0.1954	0.8843	0.1892
2	**Optimal trimmed (EW)**	**2.1574**	0.5871	2.1139	0.7887	0.9520	0.1854	0.9627	**0.2165**	0.9141	0.1767	0.9145	0.1916
2	**Optimal trimmed (OptW)**	2.1263	0.6107	2.1949	0.7697	**1.0163**	**0.2078**	0.9166	0.1943	0.9734	**0.1966**	0.8923	0.1919
5	**Single best model**	1.5782	0.4592	1.9412	0.7461	0.7732	0.1809	0.7512	0.1744	0.7878	0.1850	0.7695	0.1806
5	**Pool (EW)**	1.7750	0.4385	**2.3067**	0.7003	0.7855	0.1838	0.8472	0.2046	0.7855	0.1838	**0.8472**	**0.2046**
5	**Pool (OptW)**	0.5821	0.1171	0.6451	0.1771	0.5821	0.1171	0.6451	0.1771	0.5821	0.1171	0.6451	0.1771
5	**MCS Best model set (EW)**	1.7738	0.6519	1.9248	0.7789	0.7708	0.1827	0.7541	0.1791	0.7820	0.1863	0.7855	0.1923
5	**MCS Best model set (OptW)**	1.7815	**0.6772**	1.9182	**0.7945**	0.7663	0.1809	0.7638	0.1811	0.8091	0.1954	0.7776	0.1892
5	**Optimal trimmed (EW)**	**2.0146**	0.5871	2.0421	0.7887	0.7886	0.1854	**0.8538**	**0.2165**	0.7587	0.1767	0.7937	0.1916
5	**Optimal trimmed (OptW)**	2.0014	0.6107	2.1120	0.7697	**0.8608**	**0.2078**	0.7998	0.1943	**0.8220**	**0.1966**	0.7858	0.1919
10	**Single best model**	1.3596	0.4592	1.8233	0.7461	0.4950	0.1809	0.5068	0.1744	0.5031	0.1850	0.5287	0.1806
10	**Pool (EW)**	0.4383	0.1838	2.0841	0.7003	**1.3894**	**0.4385**	0.6022	0.2046	0.4383	0.1838	0.6022	**0.2046**
10	**Pool (OptW)**	0.4031	0.1171	0.6004	0.1771	0.4031	0.1171	0.6004	0.1771	0.4031	0.1171	0.6004	0.1771
10	**MCS Best model set (EW)**	1.6479	0.6519	1.8206	0.7789	0.5535	0.1827	0.5650	0.1791	0.5630	0.1863	**0.6099**	0.1923
10	**MCS Best model set (OptW)**	1.6662	**0.6772**	1.8196	**0.7945**	0.5443	0.1809	0.5598	0.1811	**0.5901**	0.1954	0.5998	0.1892
10	**Optimal trimmed (EW)**	1.7766	0.5871	1.9225	0.7887	0.5164	0.1854	**0.6723**	**0.2165**	0.4996	0.1767	0.5925	0.1916
10	**Optimal trimmed (OptW)**	**1.7933**	0.6107	**1.9736**	0.7697	0.6017	0.2078	0.6051	0.1943	0.5697	**0.1966**	0.6082	0.1919

Table 11: Economic implication: asset allocation with transaction cost

The table compares the out-of-sample performance of alternative methods in addressing asset pricing model uncertainty in terms of monthly certainty equivalent rates of return (CERs) expressed as a percentage and the Sharpe ratio (Sharpe) for mean-variance optimal investment strategies. The asset allocation practice uses the moments of stock return predictive distributions implied by each of the forecasting schemes to choose an optimal optimal portfolio allocation during 1978 to 2011 using three different window schemes: the full sample, rolling window and expanding window. The asset universe includes of T-bills and i) the 30 value-weighted industry portfolios or ii) the 25 size and book-to-market (B/M) portfolios. The results are presented for three different levels of risk aversion: 2, 5 and 10. The forecasting scheme includes the single best model and three types of model pool, namely, the model pool with full set of models (Pool), the model pool with trimmed model set using MCS trimming (MCS Best Model Set) and the optimal fixed trimming (Optimal Trimmed). The EW and OptW in brackets represent pooling using equal and optimal weights, respectively. This table assumes transaction costs of 50 bp per round trip. The number in bold indicates the highest economic gain in the column.

Risk aversion	**Risk aversion**	**Forecasting Schemes**	**Full Sample**				**Rolling Window**				**Expanding window**			
			30		25		30		25		30		25	
			CER	**Sharp**	**CER**	**Sharp**	**CER**	**Sharp**	**CER**	**Sharp**	**CER**	**Sharp**	**CER**	**Sharp**
2	2	**Single best model**	2.1265	0.4817	**2.2385**	0.7202	**1.1194**	0.2156	0.9651	0.1921	**1.0923**	0.2064	0.9116	0.1773
2	2	**Pool (EW)**	2.1453	0.5184	2.4470	0.7139	1.0738	**0.2163**	**1.0759**	**0.2501**	1.0738	**0.2163**	**1.0759**	**0.2501**
2	2	**Pool (OptW)**	0.6492	0.1347	0.6227	0.1772	0.6492	0.1347	0.6227	0.1772	0.6492	0.1347	0.6227	0.1772
2	2	**MCS Best model set (EW)**	1.8801	**0.7173**	2.0451	0.7505	0.8418	0.1678	0.9340	0.2073	0.8799	0.1769	0.9296	0.1984
2	2	**MCS Best model set (OptW)**	1.8679	0.6715	1.9755	0.7572	0.8310	0.1592	0.8475	0.1750	0.9012	0.1887	0.9072	0.2024
2	2	**Optimal trimmed (EW)**	2.1152	0.6202	2.0279	0.7477	1.0236	0.2102	0.8878	0.1892	0.9389	0.1812	0.8830	0.1849
2	2	**Optimal trimmed (OptW)**	**2.1706**	0.6352	2.1356	**0.7644**	0.9721	0.1914	0.9349	0.2086	0.9741	0.1960	0.9759	0.2171
5	5	**Single best model**	1.9109	0.4817	**2.1379**	0.7202	**0.9067**	**0.2156**	0.8104	0.1921	0.8728	0.2064	0.7610	0.1773
5	5	**Pool (EW)**	1.9589	0.5184	2.3183	0.7139	0.8976	0.2163	**0.9576**	**0.2501**	**0.8976**	**0.2163**	**0.9576**	**0.2501**
5	5	**Pool (OptW)**	0.6056	0.1347	0.6070	0.1772	0.6056	0.1347	0.6070	0.1772	0.6056	0.1347	0.6070	0.1772
5	5	**MCS Best model set (EW)**	1.8156	**0.7173**	1.9725	0.7505	0.7228	0.1678	0.8275	0.2073	0.7520	0.1769	0.8112	0.1984
5	5	**MCS Best model set (OptW)**	1.7952	0.6715	1.9102	**0.7572**	0.7003	0.1592	0.7402	0.1750	0.7831	0.1887	0.8076	0.2024
5	5	**Optimal trimmed (EW)**	1.9959	0.6202	1.9555	0.7477	0.8676	0.2102	0.7787	0.1892	0.7752	0.1812	0.7699	0.1849
5	5	**Optimal trimmed (OptW)**	**2.0486**	0.6352	2.0576	0.7644	0.8094	0.1914	0.8297	0.2086	0.8203	0.1960	0.8617	0.2171
10	10	**Single best model**	1.5515	0.4817	1.9703	0.7202	0.5522	0.2156	0.5526	0.1921	0.5069	0.2064	0.5101	0.1773
10	10	**Pool (EW)**	1.6483	0.5184	**2.1038**	0.7139	0.6040	**0.2163**	**0.7604**	**0.2501**	**0.6040**	**0.2163**	**0.7604**	**0.2501**
10	10	**Pool (OptW)**	0.5329	0.1347	0.5807	0.1772	0.5329	0.1347	0.5807	0.1772	0.5329	0.1347	0.5807	0.1772
10	10	**MCS Best model set (EW)**	1.7082	**0.7173**	1.8515	0.7505	0.5243	0.1678	0.6499	0.2073	0.5389	0.1769	0.6140	0.1984
10	10	**MCS Best model set (OptW)**	1.6740	0.6715	1.8014	0.7572	0.4824	0.1592	0.5614	0.1750	0.5863	0.1887	0.6416	0.2024
10	10	**Optimal trimmed (EW)**	1.7970	0.6202	1.8350	0.7477	**0.6076**	0.2102	0.5969	0.1892	0.5024	0.1812	0.5812	0.1849
10	10	**Optimal trimmed (OptW)**	**1.8453**	0.6352	1.9275	**0.7644**	0.5382	0.1914	0.6544	0.2086	0.5638	0.1960	0.6715	0.2171

Table 12: Robustness check: choice of sample split point

This table checks the robustness of the main empirical results to an alternative sample splitting point for the estimation and testing period. The table uses two different splitting points, year 10 and 25. That is, the model parameters are estimated using 10-year and 25-year rolling windows, and the forecasts are evaluated throughout the remainder of the sample periods. The table reports the log predictive score (LPSs) of the different forecasting schemes in addressing asset pricing model uncertainty. The calculation of LPSs assumes that the investor weights the accuracy of different regions of the asset return distribution as equally important. The forecasting scheme includes the single best model and three types of model pooling, namely, the model pool with full set of models (Pool), the model pool with a trimmed model set using MCS trimming (MCS Best Model Set) and the optimal fixed trimming (Optimal Trimmed). The EW and OptW in brackets represent pooling using equal and optimal weights, respectively. The table reports results for two distinct sets of test assets, namely, the 30 value-weighted industry portfolios and the standard Fama-French 25 size and B/M portfolios. The number in bold indicates the largest LPS in the column.

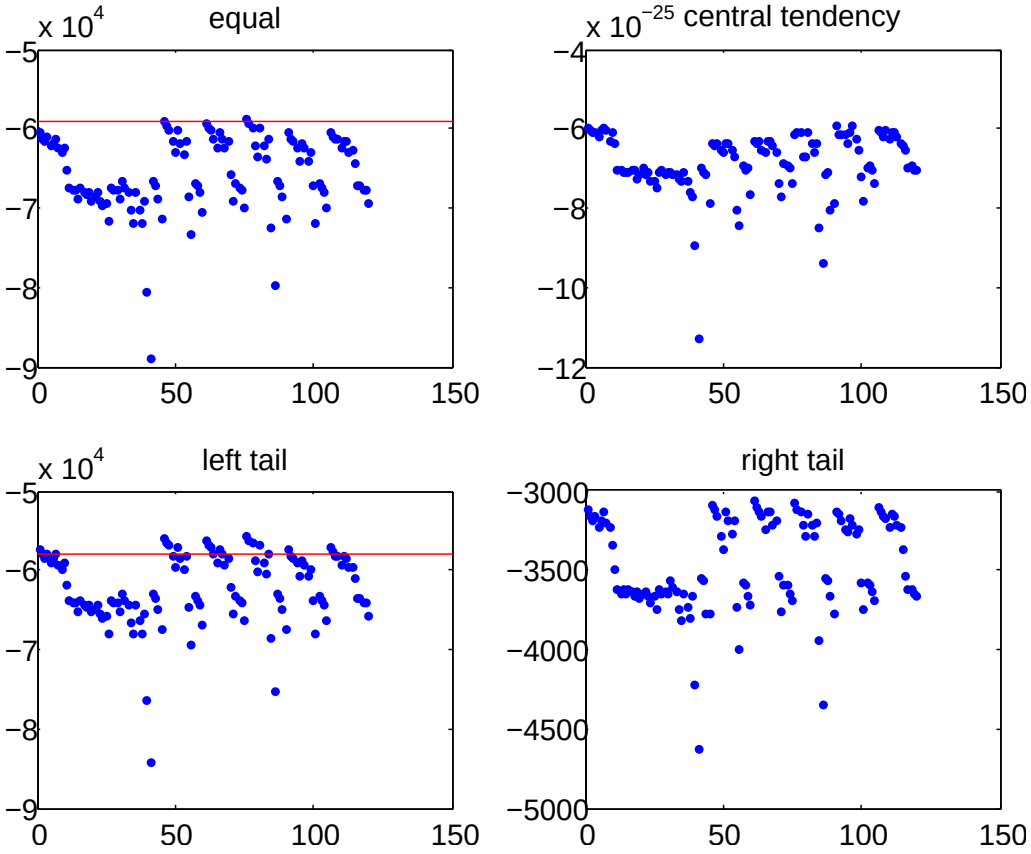
	30 industry portfolios		**25 size and B/M portfolios**	
	10 year	25 year	10 year	25 year
Single Best Model	-2578.449398	-2635.156236	1892.99653	4141.325
Pool (EW)	-2284.950213	-2180.101146	2176.15248	4594.215
Pool (OptW)	-2284.943707	-2180.099994	2505.58997	7219.102
MCS Best Model Set (EW)	-2284.950213	-2180.101146	2176.15248	4594.215
MCS Best Model Set (OptW)	-3806.520754	-4603.724938	1260.05353	2720.211
Optimal Trimmed (EW)	**-2281.2669**	-2179.537854	**2178.04098**	**4593.524**
Optimal Trimmed (OptW)	**-2281.2669**	**-2179.534645**	**2178.04098**	**4593.524**

Table 13: Robustness check: comparison of subsample periods

This table checks the robustness of the main empirical results to alternative subperiods as indicated in the column headings. The table reports the log predictive score (LPSs) of the different forecasting schemes in addressing the asset pricing model uncertainty. The calculation of LPSs assumes that the investor weights the accuracy of different regions of the asset return distribution as equally important. The forecasting scheme includes the single best model and three types of model pooling, namely, the model pool with full set of models (Pool), the model pool with trimmed model set using MCS trimming (MCS Best Model Set) and the optimal fixed trimming (Optimal Trimmed). The EW and OptW in brackets represent pooling using equal and optimal weights, respectively. Note the model weights are estimated using full-sample. The table reports results for two distinct sets of test assets, namely, the 30 value-weighted industry portfolios and the standard Fama-French 25 size and B/M portfolios in Panels A and B, respectively. The number in bold indicates the largest LPS in the column.

Panel A: 30 industry portfolios	1968-1978	1978-1988	1988-1998	1998-2008	2008-2011
Single Best Model	-12059.1693	-13278.6696	-12749.5271	-15163.6930	-5555.4678
Pool (EW)	-11904.9089	-12998.5043	-12585.1510	-14932.6278	-5361.0603
Pool (OptW)	-12329.5161	-13327.8654	-13156.0768	-15397.1549	-5504.4502
MCS Best Model Set (EW)	-11896.6035	-12958.1280	-12601.0490	-14912.9995	-5371.3520
MCS Best Model Set (OptW)	-11896.4786	-12958.0168	-12600.8625	-14912.8555	-5371.2760
Optimal Trimmed (EW)	**-11785.6696**	**-12759.2564**	**-12583.3755**	**-14824.5931**	**-5296.2185**
Optimal Trimmed (OptW)	**-11785.6696**	**-12759.2564**	**-12583.3755**	**-14824.5931**	**-5296.2185**
Panel B: 25 size and B/M portfolios					
Single Best Model	-6358.7551	-5903.7816	-5805.9099	-7583.9382	-2685.2976
Pool (EW)	-6339.1034	-5820.0233	-5798.6785	-7509.3770	-2673.4239
Pool (OptW)	-6825.6857	-6382.8183	-6297.3039	-7995.6190	-2842.3158
MCS Best Model Set (EW)	**-6197.6157**	**-5718.6465**	**-5633.4668**	**-7335.6573**	**-2601.6377**
MCS Best Model Set (OptW)	-6197.6595	-5718.6894	-5633.5409	-7335.7466	-2601.6731
Optimal Trimmed (EW)	-6310.9963	-5844.8815	-5755.5758	-7454.5168	-2645.2835
Optimal Trimmed (OptW)	-6310.9963	-5844.8815	-5755.5758	-7454.5168	-2645.2835

Panel A: 30 industry portfolios



Panel B: 25 size and B/M portfolios

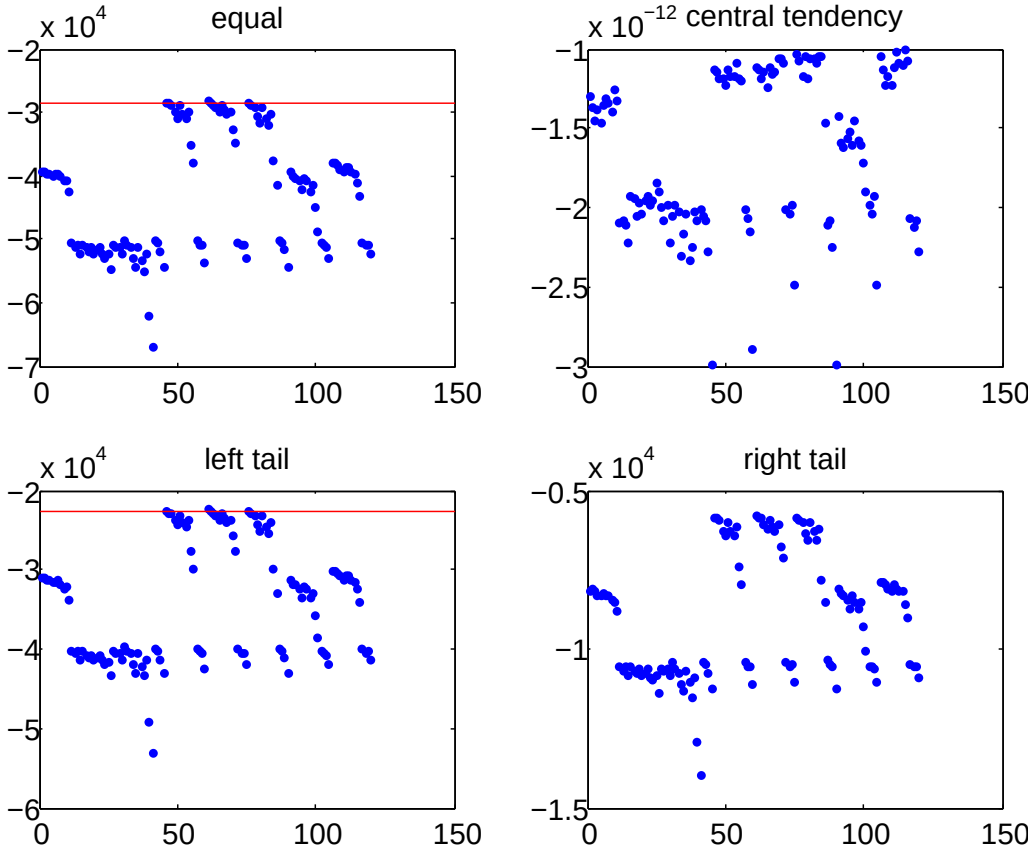
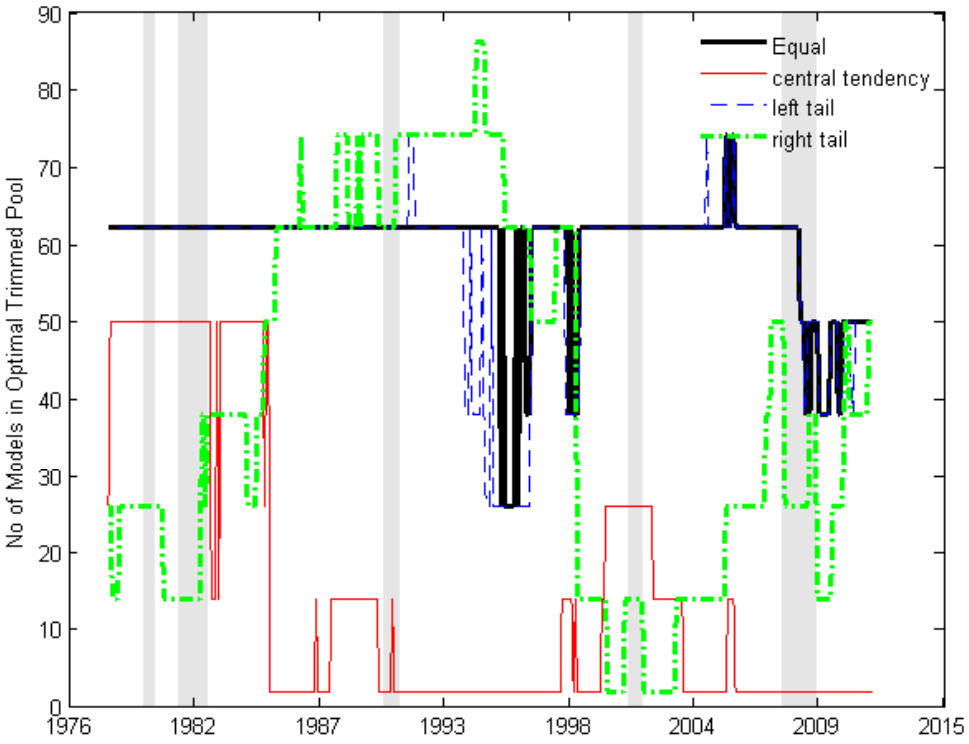


Figure 1: Comparison of individual asset pricing models: based on the Model Confidence Set (MCS) test

This figure compares and evaluates 120 asset pricing models using the MCS test developed by Hansen, Lunde, and Nason (2011). The models included are listed in AppendixA. For a given model, the WLPS is computed from predictive densities based on equation (3.11) using the full sample. The figure contains two panels illustrating the results using 30 value-weighted industry portfolios and the standard Fama-French 25 size and B/M portfolios, respectively. Each panel includes four subfigures that represent four different distributional accuracy preferences, namely, equal weighting, central tendency accuracy preference, and left and right tail preferences. The p-value of the MCS test for model comparison is computed at a significance level of 0.1%. The dots in the figure denote the individual asset pricing models. The horizontal line denotes the threshold value. The dots above the line represent the models selected into the model confidence set.

Panel A: 30 industry portfolios



Panel B: 25 size and B/M portfolios

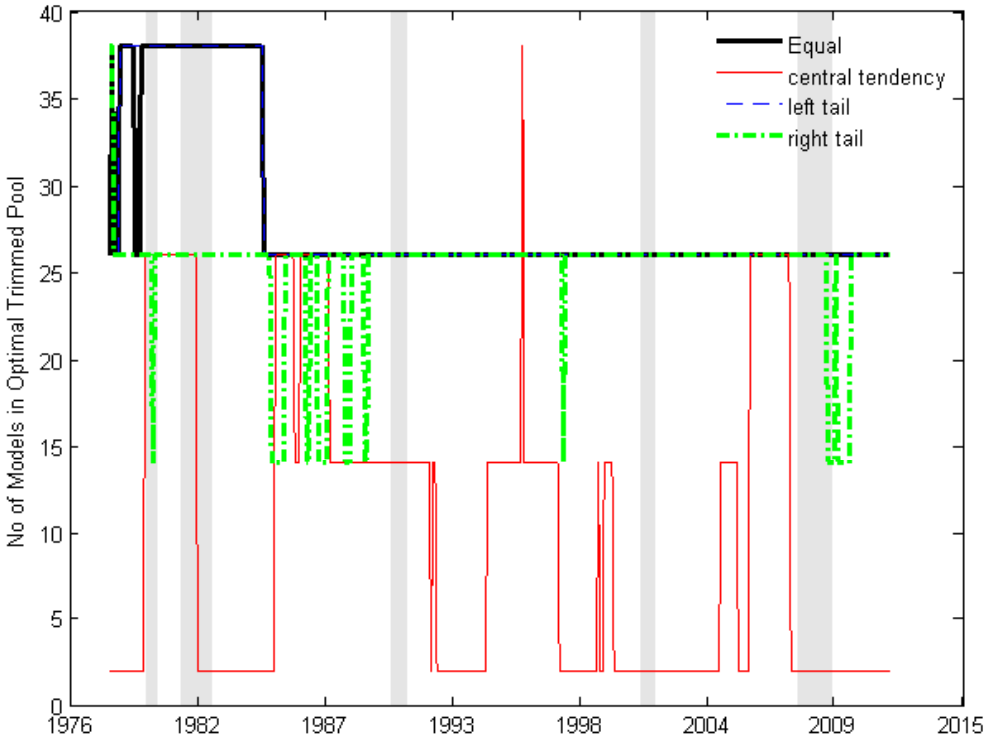


Figure 2: Time-varying number of selected models in the best forecasting scheme

This figure plots the number of models selected by the optimal trimming method over the testing period from 1978 to 2011. The forecast is based on a rolling window with 120 monthly observations. The 120 models considered are listed in AppendixA. The figure contains two panels illustrating the results using 30 value-weighted industry portfolios and the standard Fama-French 25 size and B/M portfolios, respectively. The thin solid, thick solid, dash, and dotted dash lines represent the four different distributional accuracy preferences: equal weighting, central tendency accuracy preference, and left and right tail preferences, respectively.

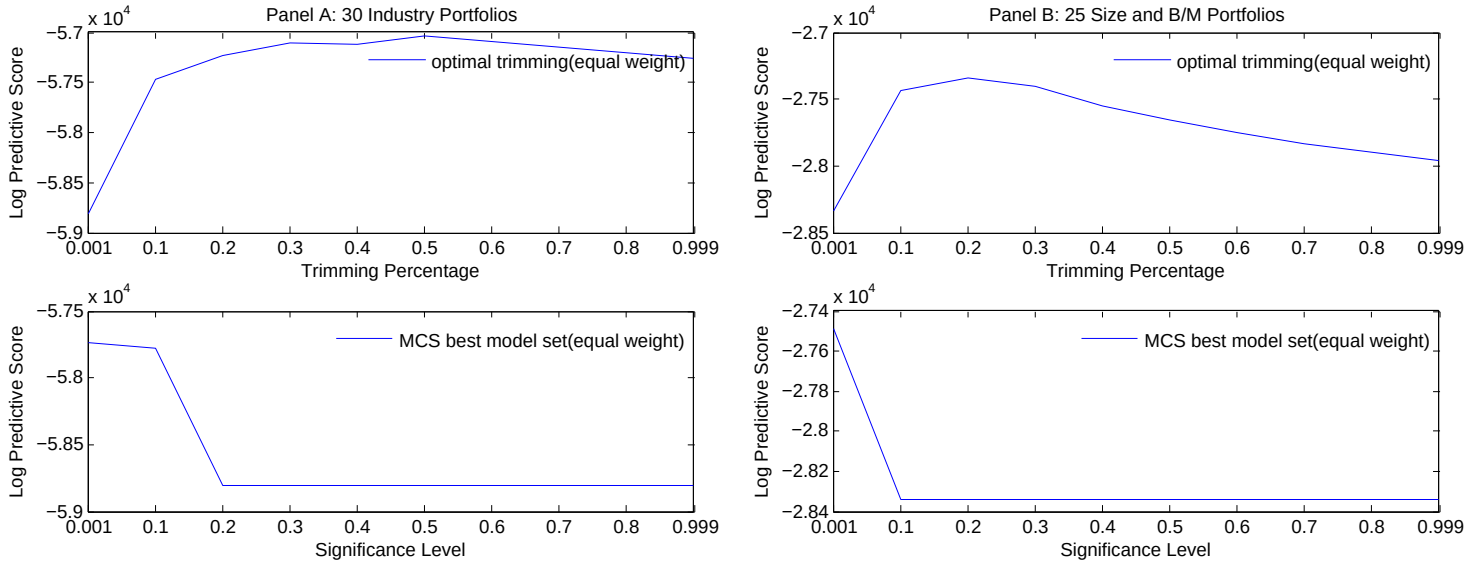


Figure 3: Sensitivity to the trimming percentage

This figure displays the log predictive score when the trimming percentage in the optimal trimming scheme varies from 0.1% to 99.9%, and the significance level in the Model Confidence Set (MSC) test (Hansen, Lunde, and Nason, 2011) ranges from 0.1% to 99.9% based on the full sample. The full set of 120 models considered is listed in AppendixA. The figure contains two panels illustrating the results using 30 value-weighted industry portfolios and the standard Fama-French 25 size and B/M portfolios, respectively.

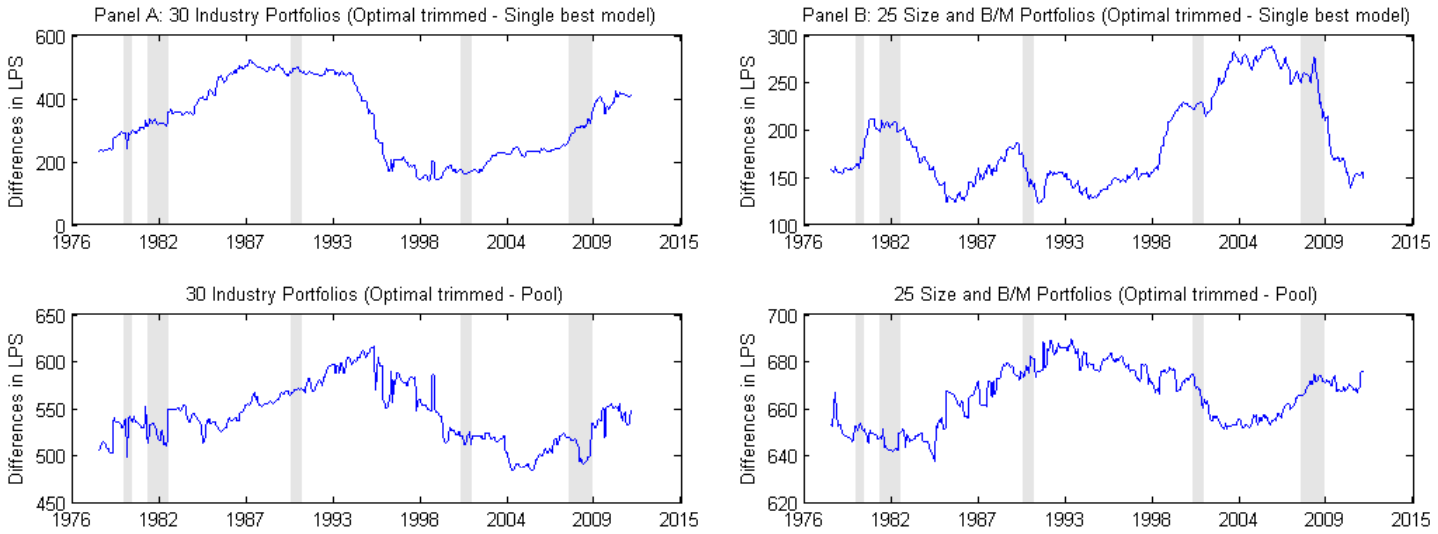


Figure 4: Differences in log predictive scores

This figure shows the difference in log predictive scores for the optimal trimming scheme and the best individual model/full model pooling. The full set of 120 individual asset pricing models considered is listed in AppendixA. The figure contains two panels illustrating the results using 30 value-weighted industry portfolios and the standard Fama-French 25 size and B/M portfolios, respectively. The top subfigure of each panel shows the difference between the optimal trimming and the best individual model, and the bottom subfigure of each panel shows the difference between the optimal trimming and the best model set constructed based on the Model Confidence Set (MSC) test (Hansen, Lunde, and Nason, 2011). The shaded areas indicate recession periods as defined by NBER.