Assessing and extending the Lee-Carter model for long-term mortality prediction

Xiaoming Liu*and Hao Yu†

Presented at the Living to 100 Symposium
Orlando, Fla.
January 5-7, 2011

*Xiaoming Liu is an assistant professor in Department of Statistical and Actuarial Sciences at University of Western Ontario, London, Ontario N6A 5B7, Canada, e-mail: xliu@stats.uwo.ca

†Hao Yu is a professor in Department of Statistical and Actuarial Sciences at University of Western Ontario, London, Ontario N6A 5B7, Canada, e-mail: hyu@stats.uwo.ca

Abstract

The prediction performance of the Lee-Carter model for long-term mortality forecast is our focus in this paper. To make a sound assessment, we set up a backtesting methodology to evaluate the prediction performance of the Lee-Carter model. We propose to use the Kolmogorov-Smirnov test to assess how close the percentile histogram resembles uniform distribution, which can complement the assessment of probabilistic prediction. We address two issues with implementing the Lee-Carter model: robustness and drift uncertainty. We propose quantile regression (QR) for robust parameter estimation of the model for time-varying index k_t. We use the bootstrap method to incorporate the drift uncertainty. Finally, we illustrate our proposed methods through examining the model performance on our simulated data as well as actual mortality data from different countries. The findings of this study suggest that the QR method improves the prediction performance of the Lee-Carter model and there exists evidence for trend changes in male mortality in the last century.

1 Introduction

The uncertainty in mortality evolution is recognized as an important risk factor for actuarial consideration that is now indispensable in mortality projection. See Pitacco (2004); Cairns *et al.* (2006); CMI (2005) for detailed review and extensive discussions on the stochastic nature of mortality improvement. This means that any mortality prediction should include both point estimates of future mortality and their associated probability distributions.

Because mortality rates tend to change at different speeds at different ages, it is often necessary to adopt a mathematical model that can describe the profile of mortality improvement in a compact format, and also allow reasonable assumptions about future development to be made and examined transparently. The Lee-Carter model (1992) seems to satisfy these criteria. The Lee-Carter model provides a simple mathematical structure that can fit the past data reasonably well, and a flexible stochastic framework that can evolve into the region of prediction as desired.

Since its publication, the Lee-Carter model has attracted great attention in literature concerning the projection of population and mortality rates. For actuarial calculation purposes, long-term mortality prediction is required. For example, the U.S. Social Security Administration (SSA) normally provides mortality forecasts for a horizon up to 90 years as a reference for the life insurance industry and retirement system. Typically, forecasted mortality rates up to 60 years or more are relevant in the evaluation of many life annuity products. Therefore the prediction performance of the Lee-Carter model for long horizon forecasts is a big concern.

It is difficult to evaluate long-term prediction performance, partially because we have to wait a long time for the history to reveal itself and partially because the random nature needs to be properly assessed with limited information. There are a few ex post studies of the Lee-Carter forecast performance based on the recent mortality experience since 1992; see Lee and Miller (2001), Bell (1997), Booth *et al.* (2002) for instance. The overall underestimation of recent mortality decline by the Lee-Carter model is consistently documented in all above reported studies. However, because these observations and the resulting conclusion are based on at most 10 new years worth of data that are still short for long-term assessment, we may wonder: Is the underestimation phenomenon the result of the underlying structural misrepresentation of

the model or just a result of chance?

To make a sound assessment, we adopt a backtesting framework that can be used to evaluate the forecast performance of the Lee-Carter model with more statistical power. The difference between backtesting and ex post examination is that, with backtesting, we can go back into the past as far as we want, given that the data are available. This is possible because the Lee-Carter model is mainly an extrapolative method based on purely statistical analysis, thus involving less subjective judgment than those that make use of expert opinions at the time of prediction. Backtesting allows us to examine how well the model would have performed overall if the Lee-Carter model were used repetitively in the past.

We examine the prediction performance of the Lee-Carter model based on both simulated data and actual mortality data from countries such as Sweden, Canada and the United Kingdom. We test the accuracy of point forecasts and density forecasts as well. We propose to use the Kolmogorov-Smirnov test to assess how close the percentile histogram resembles uniform distribution, which can complement the assessment of probabilistic prediction. To obtain robust parameter estimates in the model for the time-varying index k_t, a component of the Lee-Carter model, we propose the use of quantile regression (QR). From our simulation study, we find that drift uncertainty plays a very important role in deriving the density forecasts of future mortality. The findings of this study also suggest there exists evidence for mortality trend changes in the last century.

The outline of the paper is as follows. The Lee-Carter model is reviewed in Section 2. The QR method also is introduced there. The methods of our assessment are detailed in Section 3. Simulation studies are presented in Section 4 and applications of our proposed methodology to real data analysis in Section 5. Conclusions are made in Section 6.

2 Implementing the Lee-Carter model

2.1 Model specification

It has now been well accepted that mortality needs to be projected to allow future mortality improvement to be taken into account in the evaluation of mortality contingent products. It is also important to acknowledge that mortality trends have shown a great deal of uncertainty in

the past (Pitacco, 2004). To incorporate randomness into mortality dynamics, Lee and Carter introduced a simple yet powerful statistical model for fitting and projecting mortality. Under the Lee-Carter framework, the log of age-specific central mortality rates are described as

$$\log m_{xt} = a_x + b_x\, k_t + \epsilon_{xt}\,, \tag{1}$$

where $x = 1, 2, \ldots, n$ represent ages and $t = 1, 2, \ldots, t_0$ represent years. Hence, through the Lee-Carter decomposition, the mortality improvement over time can be summarized with two age factors a_x and b_x, and one time-varying index k_t. Here k_t represents the time series of the general level of mortality, while a_x describes the age profile averaged over time, and b_x determines how much, at each age, the mortality rate responds to the changes in k_t.

After the Lee-Carter model is fit to a selected data set, a_x's and b_x's are treated as constants and the values of k_t are modeled by a time series. In the original paper of Lee and Carter (1992), it is suggested that an autoregressive integrated moving average, specifically ARIMA(0,1,0), is the most appropriate model for k_t, even though in some cases other ARIMA models might be preferable. The ARIMA(0,1,0) is equivalent to a random walk with drift and can be written as follows:

$$k_t = k_{t-1} + c_1 + \xi_t, \tag{2}$$

where ξ_t is $N(0, \sigma^2)$, independently and identically distributed. The drift term of this random walk, c_1, and its standard deviation, σ, are estimated from $k_1, k_2, \ldots, k_{t_0}$. Forecasts of future values of k_t $(k_{t_0+1}, k_{t_0+2}, \ldots)$ can then be recursively generated using formula (2). More specifically, to forecast the time-varying index at time $t_0 + n$ given the data available up to t_0, the following equation is used:

$$k_{t_0+n} = k_{t_0} + n \cdot c_1 + \sum_{j=1}^{n} \xi_j\,. \tag{3}$$

2.2 Density forecasts of mortality rates and life expectancy

The projected $k_t\,(t > t_0)$ can then be substituted into formula (1), together with the estimated a_x and b_x, to calculate the forecasted age specific mortality rates in all future years. Assume the force of mortality μ_{xt} is constant within each age band. Denote as $p_x(t)$ the survival probability that an individual aged x in year t reaches age $x + 1$, as $q_x(t) = 1 - p_x(t) = 1 - e^{-\mu_{xt}}$ the

corresponding death probability. Consider $e_0(t)$, the life expectancy[1] at birth of newborns in year t, which can be calculated by

$$\begin{aligned} e_0(t) &= \frac{1}{2} + \sum_{n=1}^{\infty} \prod_{x=0}^{n-1} p_x(t)\,, \\ &= \frac{1}{2} + \sum_{n=1}^{\infty} e^{-\sum_{x=0}^{n-1} \exp(a_x + b_x\, k_t)}\,. \end{aligned} \tag{4}$$

Formula (4) shows that the life expectancy $e_0(t)$ is a function of the predicted time varying random variable k_t. This means the stochastic nature of the time series k_t is passed to all variables derived from mortality rates, including life expectancies. The resulting distribution for $e_0(t)$ can thus provide a probabilistic description of the variable in the future year t. The probability density forecast contains not only the "best" point forecast of our variable of interest, but also the information on how likely the variable may differ from the point forecast. In this paper, we are interested in using $e_0(t)$ as a summary variable to check the prediction performance of the Lee-Carter model. We will examine the errors in its point forecasts and the validity of its density forecasts.

2.3 Non-robustness of least squares and a solution: quantile regression

Clearly, the accuracy of the model for k_t determines the quality of mortality forecasts. In the original Lee and Carter paper, the drift term for k_t (in model (2)) is estimated using the method of ordinary least squares (LS). That is, the parameter c_1 is chosen to minimize the sum of squared errors:

$$\min_{c_1 \in R} \sum_{t=1}^{t_0} (\Delta k_t - c_1)^2,$$

where $\Delta k_t = k_t - k_{t-1}$. This leads to

$$\hat{c}_1 = \frac{k_{t_0} - k_0}{t_0}\,, \tag{5}$$

[1]Note that the calculation for $e_0(t)$ is conducted in the so-called "period" direction because we are interested in using $e_0(t)$ as a kind of summary variable for the mortality profile of the corresponding year. For other purposes such as the evaluation of a life annuity contract, it is suggested that the "cohort" direction should be used. For more details regarding the "period" or "cohort" distinction, see Pitacco (2004).

where k_0 and k_{t_0} are the first and last values of the time-varying index for the base period. Obviously, the estimate $\hat{c}_1$ is very sensitive to the first and last years of the data and not robust against outliers and extreme abnormal values.

In this section, we introduce a robust estimation procedure for the drift term of k_t, where the parameter c_1 in model (2) is obtained by performing the optimizition

$$\min_{c_1 \in R} \sum_{t=1}^{t_0} |\, \Delta k_t - c_1 |\,. \tag{6}$$

The method we use in (6) is median regression, minimizing the sum of absolute deviations. Median regression is a special case of quantile regression first introduced by Koenker and Bassett (1978). The review paper by Koenker and Hallock (2001) outlines more recent developments on the QR method. QR leads to estimates of specific quantiles of the response variable, in contrast to LS which provides estimates that approximate the conditional mean of the response variable. In particular, median regression results in estimates of the median response. QR is robust in response to large outliers by comparison with the LS method; outliers can seriously distort LS parameter estimates, while QR is resistant to the effects of outliers. The motivation for the QR method comes from the need to properly handle shocks due to historic events such as the 1918 Spanish flu epidemic and the 2003 SARS outbreak.

3 Evaluation methods for mortality forecasts

Since Lee and Carter (1992), many modifications and extensions to the Lee-Carter model have been developed. Among those, Carter (1996) has considered a state space model for k_t in which the drift term is itself a random walk; Renshaw and Haberman (2003) have added a second or even third bilinear term to the Lee-Carter model; Renshaw and Haberman (2006) have introduced the cohort period effect. In general, those modifications improve the fit of the model to some specific data. However, as shown in Dowd *et al.* (2008) and Cairns *et al.* (2009), those models may generate implausible predictions because the models are too specific or complicated, and parameter estimates are too sensitive. Therefore, in this paper, we don't attempt to test the Lee-Carter model against other alternatives, or the random walk with drift assumption for k_t against other alternatives. We have assumed that a random walk with drift

is the forecasting model for the time-varying index k_t always used. We are interested in testing how well the Lee-Carter model would have performed overall if it were used repetitively in the past. We are also interested in examining the robustness of Lee-Carter model prediction when the data are subjected to occasional large shocks to human mortality, such as the 1918 Spanish flu pandemic.

In this section, we describe the evaluation methods used to assess forecast performance of the Lee-Carter model. Because the stochastic mortality modeling approach provides point forecasts and density forecasts as well, we assess the prediction performance in both ways. We consider the forecasts within a 30 year horizon as short-term forecasts and beyond a 30 year horizon as long-term forecasts. Instead of using mortality rates at any specific age as a benchmark, we use life expectancy at birth e_0 as an overall measure for the age-specific mortality pattern of each year.

3.1 Backtesting Framework

Following Dowd *et al.* (2008), we refer to our evaluation framework as a backtesting method. Here is what we are going to do in backtesting.

(i) We first select a *base period* which is used to estimate the parameters a_x, b_x and k_t in equation (1). In this step, we adopt the maximum likelihood estimation (MLE) approach, first proposed by Brouhns *et al.* (2002).

(ii) Then the time-varying index k_t is fit to ARIMA(0,1,0) by either *the LS method* as in original Lee-Carter paper or *the QR method* as described in the previous section.

(iii) Forecasting starts from the last year of the base period, which is normally referred to as a "*jump off*" year. With each given base period and fitted Lee-Carter model, we can derive forecasted mortality rates based on the forecasted values of k_t, together with estimates a_x and b_x. For a selected *forecast window*, we are able to obtain sequential values of forecasted life expectancy $\hat{e}_0(t)$ corresponding to different horizons within that window. We are also able to derive the probability density distribution for each $e_0(t)$. In this paper, because we are interested in long-term forecast performance, our forecast horizon is up to 60 years whenever possible, based on data availability.

(iv) Finally, we compare our forecasts for $e_0(t)$ with the actual outcomes to obtain various forecast performance measures that are provided later in this section.

It is worth remarking that the sequential forecasts $\hat{e}_0(t)$ comprise different horizon forecasts. Generally, longer horizons lead to less accurate forecasts. We should only compare the accuracy of forecasts with similar horizons. To make a more formal statistical test, we need to construct a sample of forecasts for each horizon. In the simulation study described later, we generate 10,000 scenarios for each simulation scheme to make the assessment. The purpose is to check the overall model performance under various situations.

In the analysis of real data, we perform a rolling procedure using fixed length base periods over historical data as the jump off year moves forward through time. Each time, we refit the model to the chosen base period of data, and obtain the forecasts correspondingly. In this way we can check the overall dynamic prediction performance of the Lee-Carter model with more statistical power.

We evaluate the goodness of forecasts by implementing the assessments described in the following sections.

3.2 Forecast error criteria

We have considered various error measures to examine the accuracy of the forecasts for $e_0(t)$, including the root mean square error (RMSE), the mean absolute percent error (MAPE), the average error (bias), and the proportion of actual values that fall within the 95 percent probability interval (coverage). The better forecasts should come up with smaller values of RMSE, MAPE and bias, while the actual coverages should be close to the nominal one, e.g., 95 percent, as used in this paper. These measures provide criteria for model performance in terms of best point forecasts as well as a probabilistic profile.

3.3 Percentile histogram and Kolmogorov-Smirnov test

In simulation and real data analysis, percentile values of the real $e_0(t)$ in the probability distribution can be used to check the adequacy of probabilistic prediction. If prediction is correct, the frequency distribution of the percentiles should closely resemble a standard uniform distri-

bution. A common way to check this property is to draw percentile histograms as in Lee and Miller (2001). Percentile histograms can also provide intuition into the relationship between the outcomes and their associated distributions. If the plot is more clustered in the center than in the tails, it indicates that the variation of the variable is overpredicted, and vice versa. When more values fall on the right side, then we are underestimating the true value most of the time.

To make comparisons, a statistical test of how close percentiles resemble the uniform distribution is desirable. In the following we briefly describe the idea of using the Kolmogorov-Smirnov test to evaluate the percentiles. A more detailed treatment on the Kolmogorov-Smirnov statistic and its relevant results is given in Appendix A.

Let $\{p_1, \ldots, p_n\}$ be the independent and identically distributed (iid) sample percentiles at a specific forecast horizon and $\hat{F}_n(s; p_1, \ldots, p_n)$ be the empirical cumulative distribution function (CDF) function based on the data $p_1, \ldots, p_n$, i.e.,

$$\hat{F}_n(s; p_1, \ldots, p_n) = \frac{1}{n} \sum_{i=1}^{n} I(p_i \leq s),$$

where $I(\cdot)$ is an indicator function. Then we can define the empirical process $B_n(s)$ as

$$B_n(s) = \sqrt{n}\,(\hat{F}_n(s; p_1, \ldots, p_n) - s). \tag{7}$$

Notice that s in formula (7) is the CDF of the standard uniform distribution. In Proposition A.1 of Appendix A, we have shown that $\{p_1, \ldots, p_n\}$ is a sample from a standard uniform distribution asymptotically. With the help of the strong law of large numbers, we can show that, almost surely,

$$\sup_{0 \leq s \leq 1} |\hat{F}_n(s; p_1, \ldots, p_n) - s| \to 0.$$

This justifies the idea of comparing the empirical CDF of sample percentiles to the standard uniform distribution.

Now define the Kolmogorov-Smirnov statistic as

$$ks.statistic = \sup_{0 \leq s \leq 1} |B_n(s)|. \tag{8}$$

In Proposition A.3 and Corollary A.4 of Appendix A, it is shown

$$\sup_{0 \leq s \leq 1} |B_n(s)| \to \sup_{0 \leq s \leq 1} |B(s)|, \quad \text{in distribution,}$$

where $\{B(s)\}$ is a standard Brownian bridge. This result gives the asymptotic distribution of the *ks.statistic* when $p_1, \ldots, p_n$ are iid sample points. The critical value at 95 percent confidence level for Kolmogorov-Smirnov test is 1.36, which can be used to determine whether to accept or reject the null hypothesis at the 5 percent level.

In the context of dynamic prediction of the Lee-Carter model, the iid condition may be violated. In this case, it can be shown that under the null hypothesis that $\{p_1, \ldots, p_n\}$ follows the standard uniform distribution, the asymptotic distribution of the *ks.statistic* exists, but its actual form depends on the correlation structure of $\{p_1, \ldots, p_n\}$. Hence, when the iid condition does not apply, the inference based on the Kolmogorov-Smirnov test may not be accurate. But, the value of *ks.statistic* can still be used as a guideline to rank the goodness of forecasts of density distributions using different methods.

4 Simulation study

In applying the Lee-Carter model for long-term mortality prediction, two types of problems need to be distinguished: one is the drift in the values of k_t; the other is the structural change in model components a_x, b_x and k_t. Drift is a problem related to the effectiveness of model estimation, while the structural change is a problem of model specification. Oftentimes, we feel that the two problems are confounded together and this makes it difficult to diagnose. The complete solution to this issue is beyond the scope of this paper. What we have attempted in the paper is to isolate the problem of model effectiveness by using simulated data and then to assess model prediction performance by using real data. Comparing the behavior of model performance over different data sets (simulated and real data) allows us to detect potential structural change in the real data to some extent.

We address two issues in our simulation study. First, we investigate the importance of taking account of drift uncertainty in deriving the density forecasts based on the Lee-Carter model. Secondly, we would like to see if quantile regression (QR) can improve robustness of parameter estimation of the Lee-Carter model, thus resulting in more reliable forecasts overall. In particular, we examine the model behavior when the underlying data is subject to irregular jumps, mimicking the situation when catastrophic or pandemic events happen. Therefore, we

generate two cases of data for each concerned population.

- Case 1 represents the situation where the underlying data truly follows the Lee-Carter model;
- Case 2 represents the situation where the underlying data follows the Lee-Carter model most of the time but the innovations in the dynamics of time-varying index k_t are subject to rare irregular large shocks.

4.1 Simulation scheme

In this part, we provide steps used to generate underlying raw data. The steps are applied to both cases 1 and 2 unless they are mentioned specifically.

Step 1. Start with given data D_{xt} and ETR_{xt} from a specific population with $t_0 + n$ consecutive years, where D_{xt} is the number of deaths at aged x in year t and ETR_{xt} is the exposure-to-risk or the person-years lived by people aged x in year t. Later, we will use the first t_0 ($= 40$ in this paper) years as the base period and next n years as the forecast window. At this step, the MLE procedure of the Lee-Carter model is applied to the whole period. This provides us a fixed set of $\{a_x, b_x, k_t\}$ as ***seed*** to generate the raw data that follow Lee-Carter model.

Step 2. Compute the sample mean and standard deviation for the difference of k_t, denoted as k_{mean} and k_{sd}.

Step 3. Generate a random sample ξ_t with sample size $t_0 + n$, mean 0 and standard deviation k_{sd}. For Case 1, a normal sample is generated. As for Case 2, the sample to be generated contains 95 percent of the sample points following a normal distribution with standard deviation $k_{sd}/\sqrt{2.2}$ and 5 percent of the sample points following a normal distribution with standard deviation five times as large as the previous ones. In other words, ξ_t is a mixture of normal random variables with the following CDF:

$$F_\xi(x) = 0.95\,\Phi(x; 0, k_{sd}^2/2.2) + 0.05\,\Phi(x; 0, (5\,k_{sd})^2/2.2), \tag{9}$$

where $\Phi(x; \mu, \sigma^2)$ is the CDF of a Normal $N(\mu, \sigma^2)$ r.v., and 2.2 is a scale factor[2] so that the

[2]In our simulation study, we use a scale factor 2.2 to ensure that the random shocks of the model have the

standard deviation of the mixture normal is exactly k_{sd}. Case 2 represents the situation where irregular large shocks may happen occasionally.

Step 4. Simulate raw Lee-Carter data as follows. First, a sample path of k_t for $t_0 + n$ years is computed as

$$k_{t,\,path} = k_0 + k_{mean}\, t + \sum_{j=1}^{t} \xi_j.$$

Then this path $k_{t,\,path}$, together with estimates a_x and b_x (from Step 1) and ETR_{xt}, generates a new set of deaths D^*_{xt} that follow a Poisson distribution $D^*_{xt} \sim \text{Poisson}(\lambda_{xt})$ with $\lambda_{xt} = ERT_{xt} \exp(a_x + b_x\, k_{t,\,path})$.

This gives us raw Lee-Carter data $(D^*_{xt}\,,\, ETR_{xt})$ from the seed $\{a_x, b_x, k_t\}$.

Step 5. Now we apply the backtesting procedure (i) to (iv) described in Section 3 to the data $(D^*_{xt}\,,\, ETR_{xt})$. We use the first t_0 years as the base period to find $a_{x,sim}$, $b_{x,sim}$, and $k_{t,sim}$ which are then used to predict mortality rates in the chosen forecast window.

Here, we need to give some details on how to use the bootstrap method to derive forecasted distribution for variable $e_0(t)$. The basic idea is to generate enough future samples so that the sample distribution can approximate the true one. We consider two types of mortality forecasts.

(a) Assume that the drift term in the model for k_t is known with certainty[3] – a case referred to as "drift certain" (DC)

In this case, we only need to incorporate the innovations from k_t in generating the future path. A bootstrap method is conducted by resampling a set of errors $\{\xi^*_t\}$, $t = t_0 + 1, \cdots, t_0 + n$, from

$$\{\xi_{t,sim} = k_{t,sim} - k_{t-1,sim}\}.$$

ξ^*_t provide a representation of variations in the future path of k_t that lead the realized outcomes to differ from the forecasted ones.

same standard deviation for Case 1 and Case 2. An alternative simulation design such that the regular random shocks have the same standard deviation as in Case 1, but the irregular ones have a bigger standard deviation is also of interest and can easily be obtained from modifying (9).

[3]Denuit (2008) provides an analytic method to derive the quantiles and distribution function for $e_0(t)$ in any future years using comonotonic approximation under this assumption.

(b) Assume that the drift term in the model for k_t is only estimated, thus the "true" value is unknown[4] – a case referred to as "drift uncertain" (DU).

We think this is a more plausible case. To incorporate the forecast error due to the estimate of the drift term in k_t, we use a bootstrap to create uncertainty for the drift term. For given $k_{t,sim}$, a sample of perturbations with size t_0 is generated from $\{\xi_{t,sim}\}$, now denoted as $\xi^*_{t,sim}$, $t = 1, \cdots, t_0$. Combined with $k_{mean.sim}$, the sample mean of $\xi_{t,sim}$, a new $k^*_{t,sim}$ is generated as

$$k^*_{t,sim} = k_{0,sim} + k_{mean.sim}\, t + \sum_{j=1}^{t} \xi^*_{j,sim}.$$

Future innovations in the path of k_t in case DU is generated from $k^*_{t,sim}$ as in (a). The future mortality rates and life expectancy $e_0(t)$ are then calculated respectively.

Mortality forecasts (a) or (b) will be repeated many times until the density or quantiles of predicted values can be computed reliably. In our simulation, we choose 10,000 as the total bootstrap number to generate "simulated" forecast distribution for $e_0(t)$ for each raw data set. Sample medians are used as forecasted values $\hat{e}_0(t)$, compared with its corresponding "true" values from Step 4. As a result, we can derive RMSE, bias, etc. We also record if the 95 percent prediction interval at each horizon covers the true value with this data set.

Step 6. Repeat step 3 to 5 many times to generate Lee-Carter model scenarios. In this paper, 10,000 scenarios are generated. Overall we have conducted 10,000 by 10,000 simulations for each simulation scheme. This requires lengthy simulation times that could take up to half year on a typical personal computer. To speed up, we use the parallel computing package Rmpi in R (Yu, 2002).

4.2 Density forecasts with drift uncertainty

It is clear that the estimation of the drift term is central to the forecasts. Due to the non-stationary nature of the random walk and the fact that we only have limited amount of data, the drift estimate must inevitably be subject to a degree of uncertainty. To properly reflect

[4]In this paper, other types of parameter uncertainty are ignored. One reason is because the parameter uncertainty from a_x and b_x is relatively minor (see Liu and Braun, 2010, for details).

this uncertainty in the forecasts, we adapt the bootstrap method to generate the forecast distribution of $e_0(t)$. Details have been given in Step 5 in the previous subsection. In this part, we examine the effects of incorporating drift uncertainty in the mortality prediction.

Figures 1 to 3 show the simulation results using the seed from Swedish female mortality rates from 1907 to 2006. All data used in this paper are downloaded from www.mortality.org. Figure 1 presents various forecast errors: RMSE, MAPE, and average bias. Figure 2 gives the statistics related to assess the adequacy of density forecasts: Kolmogorov-Smirnov statistics, coverage, and average confidence interval width. In all plots, we compare the results of involving drift uncertainty (labeled as DU) or not involving (labeled as DC), for Cases 1 and 2 respectively. The left panel of the figure shows the results based on Case 1, the right one on Case 2. Corresponding simulation results using the seed from Swedish male mortality at the same period are given in figures 13 to 15 in Appendix B. The estimation and prediction methods used in figures 1 to 3 and figures 13 to 15 are the LS method.

We observe from this study that

- In terms of forecast errors, there is no clear advantage with DU, though many times but not always we obtain smaller forecast errors with DU.

- In terms of prediction distribution, Kolmogorov-Smirnov statistic for DC density becomes larger as forecast horizon goes longer. However, Kolmogorov-Smirnov statistic for DU density is stable and stays around its critical value 1.36 for Case 1 over all horizons. For Case 2, although the value of Kolmogorov-Smirnov statistic for DU density is bigger than 1.36, it is significantly smaller than its counterpart of DC density.

- The goodness of forecasts for density can be further examined through coverage plots (the second row panel in Figure 2). The results are consistent with Kolmogorov-Smirnov statistics, showing that the coverage with DU in Case 1 is very close to the nominal one, e.g. 95 percent, while the coverage with DC drops in a linear pattern from 93 percent to about 75 percent. Therefore, taking account of drift uncertainty can provide remarkable effects for the long term distribution prediction. Otherwise, the prediction confidence intervals will be too narrow to reflect the true risk in mortality. See the plot of average confidence interval widths as in the third row panel of Figure 2).

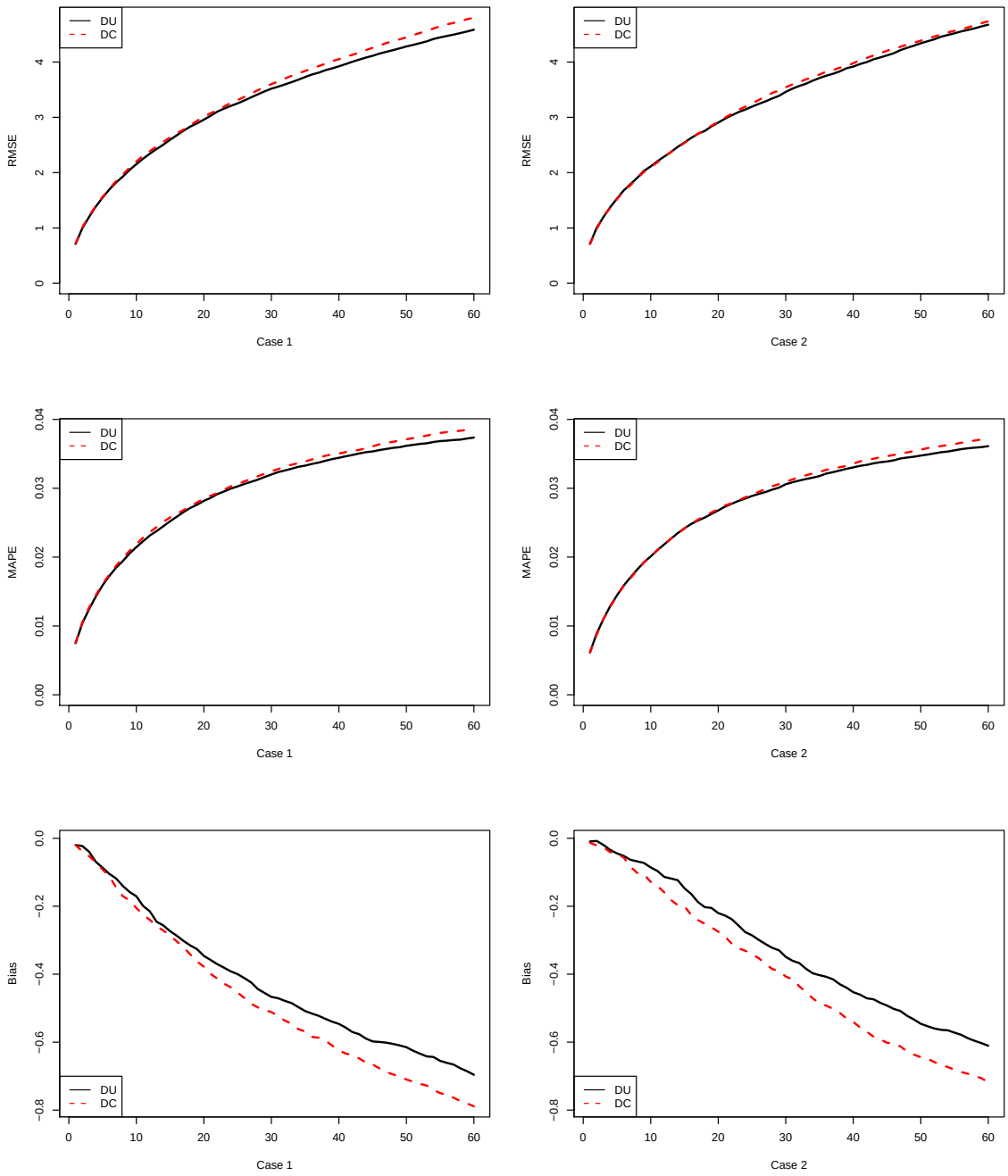

Figure 1: Simulation study over the effect of drift uncertainty on forecast performance. The seed is obtained from *Swedish female* mortality data from 1907 to 2006. Forecast errors of RMSE, MAPE, and bias for $e_0(t)$ for 60-year horizons are presented for both Case 1 (left column panel) and Case 2 (right column panel).

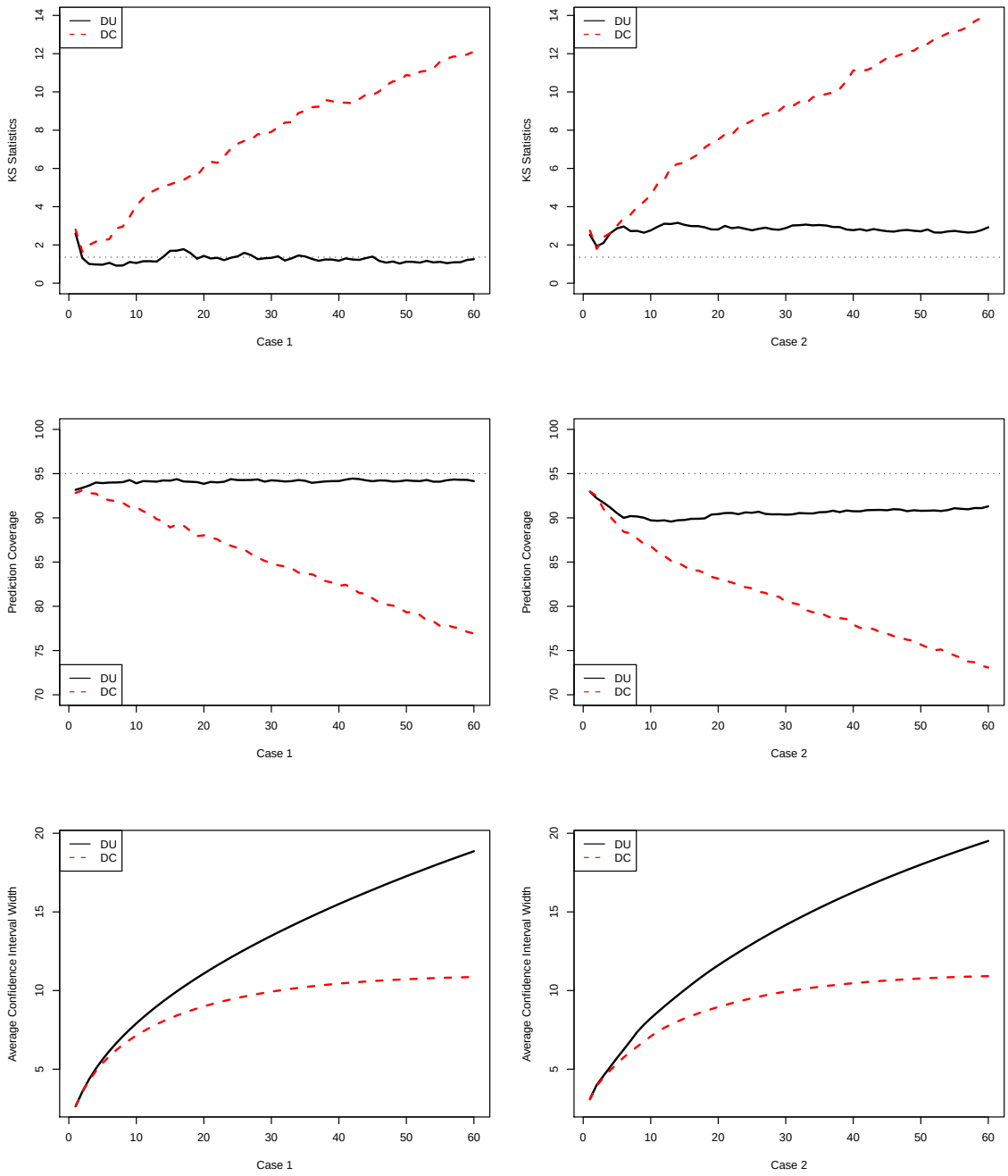

Figure 2: Simulation study over the effect of drift uncertainty on forecast performance. The seed is obtained from *Swedish female* mortality data from 1907 to 2006. Density forecast for $e_0(t)$ for 60-year horizons are assessed by Kolmogorov-Smirnov statistics, coverage and average confidence interval width for both Case 1 (left column panel) and Case 2 (right column panel). The critical value 1.36 from Kolmogorov-Smirnov test and the nominal coverage 95 percent are also marked.

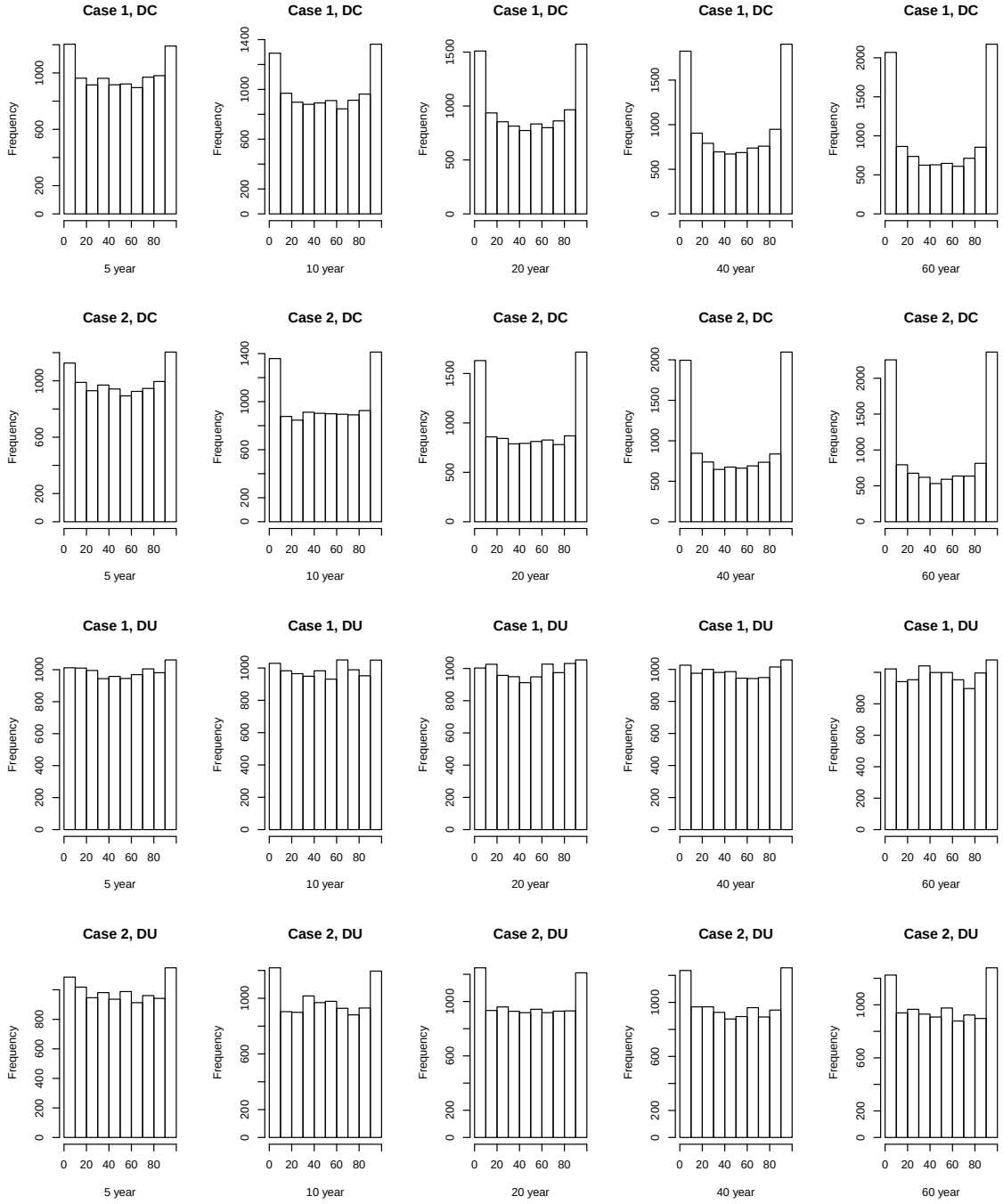

Figure 3: Simulation study over the effect of drift uncertainty on forecast performance. The seed is obtained from *Swedish female* mortality data from 1907 to 2006. Percentile histograms for forecasts $e_0(t)$ for five-, 10-, 20-, 40- and 60-year horizons are presented for Case 1 (DC), Case 2 (DC), Case 1 (DU) and Case 2 (DU), respectively.

- Percentile histograms in Figure 3 reveal that the actual distribution of percentiles from DC tends toward the two ends 0 and 100 as seen in the first two row panels. The longer the forecast horizon is, the heavier the tails of the histogram are. This indicates that the forecasts underpredict the degree of uncertainty for variable $e_0(t)$. In other words, the true value is likely to fall outside of 95 percent prediction intervals more than 5 percent of the time if the density forecasts show this kind of pattern. However, with a DU density (the last two row panels of the figure), the percentile distribution is very close to uniform between 0 and 100 as expected.

- For all measures discussed above, Case 1 displays better forecast performance than Case 2. This is because the LS method is less efficient in Case 2, where the underlying data contain irregular shocks.

So far, we have shown it is necessary to take account of drift uncertainty in deriving probabilistic forecasts. Therefore, from now on, we will only present the forecasts involving drift uncertainty.

4.3 Quantile regression (QR) vs. least squares (LS)

In Section 4.2, all of the results are based on the LS method. We have seen that the LS method with DU produces reasonably good forecasts for Case 1. But with Case 2, a lack of robustness is evident in dealing with data variations. We expect that QR can perform better in this kind of situation, where the underlying structure is subject to some irregular large shocks. More specifically, we examine the forecast results using either LS as in the original Lee and Carter paper or QR as proposed in this paper to fit and predict the time-varying index k_t in step [iii] of the backtesting procedure, while the rest of the steps remain the same. We present the comparison in this section.

Figures 4 to 7 give the results of implementing LS or QR with simulated data, using the seed from the Swedish female mortality data as in the previous section. Figure 4 contains various forecast error measures. Figure 5 gives the assessments related to density forecasts. As before, the left panel shows the results based on Case 1, and the right one on Case 2. The dashed lines represent the results with LS, whereas the solid lines are from QR. We also plot the results

using only 20-year as a base period to do prediction in those figures. Red lines are for the 20-year base period and black lines are for the 40-year base period. Figures 6 and 7 show the percentile histograms for cases 1 and 2 respectively.

Let's first focus on the 40-year base period generated results (black lines in figures 4 and 5). Here are some remarks.

- In terms of forecast errors RMSE, MAPE and bias, LS performs better in Case 1, and QR performs better in Case 2. It makes sense that LS outperforms QR in Case 1 because LS is supposed to be optimal if the model and the underlying structure match.

- In terms of the probabilistic aspect of the prediction (see Figure 5), we see that for Case 1, the values of *ks.statistic* from LS are mostly smaller than the critical value 1.36, indicating that we can not reject the null hypothesis—sample percentiles $\{p_1, \ldots, p_n\}$ follow a standard uniform distribution. As a result, it shows the predicted density distributions are adequate for Case 1. In addition, the *ks.statistic* from QR follows the track of the *ks.statistic* from LS closely, with some improvement for at least 40 years. LS coverage, though a little bit smaller than 95 percent, is stable and consistent. On the other hand, QR coverage, which is also stable, is a bit higher than 95 percent—a sign of overcorrection. As a result, the prediction intervals from QR are wider than from LS.

  For Case 2, the results are reversed. QR generates lower *ks.statistic* values and higher coverage than LS, showing a robustness property while the underlying data deviates from the assumption.

- It is remarkable to notice that for Case 2, the coverage from QR is higher than the coverage from LS, and this is usually achieved with narrower confidence intervals from QR. This again shows the robustness of QR with data variations and limited amount of data.

- The percentile histograms in figures 6 and 7 show that the density forecasts perform reasonably well with QR, even for forecast horizons of 60 years. But with LS, the histogram tends toward the two ends, indicating a sign of underestimation in mortality uncertainty. They further confirm our aforementioned conclusion based on the Kolmogorov-Smirnov test and coverage.

We have also examined the results for other types of data, i.e. using seeds from different countries and genders, we consistently see an improvement in forecasts by using QR in Case 2. Additional results are provided in Appendix B, including the simulation study based on Swedish male mortality data from 1907 to 2006, Canadian female and male mortality data from 1921 to 2006, and U.K. female and male mortality data from 1922 to 2006. We have shorter forecast horizons for Canadian and U.K. mortality due to data availability.

Other points to note are:

- With the results based on the 20-year base period, the major conclusion about LS and QR remains. With Canadian and U.K. mortality data (in Appendix B), the results from the 20-year base period are generally worse than their 40-year base counterpart, indicating a sign of instability. This puts a caution to users of the Lee-Carter model. In applications, market practitioners tend to use short base period (for example, 10 years) to calibrate the model because the linear pattern assumption for the time-varying index k_t might fit the data better over a shorter period. We find this is dangerous because a well fitted model doesn't guarantee good predictions. Similar phenomenon has also been found in Cairns *et al.* (2009)

  However, we also notice this observation doesn't apply to the results from Swedish mortality data, both female and male. This is probably because the seeds for 40-year base predictions are derived from the mortality data containing year 1918 when Spanish flu took place, thus the overall uncertainty in this simulation study is bigger than the other circumstances.

- In all of our simulations, we find a systematic problem when implementing the Lee-Carter model. That is, the bias—calculated by taking the average of (predicted value - actual value) over the same horizon forecasts—tends to be negative regardless of which methods and data we use. We can not explain what causes this problem.

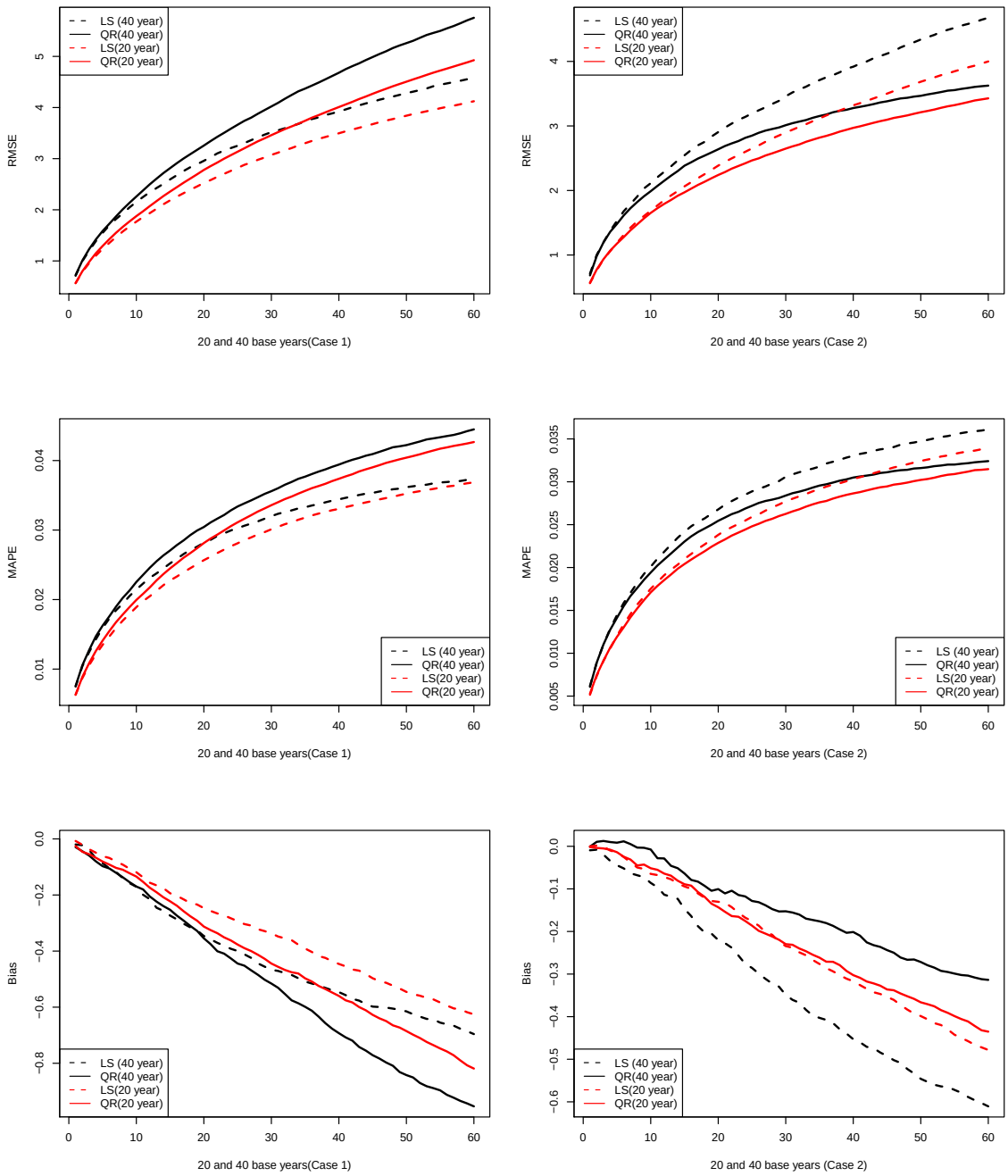

Figure 4: Simulation study on forecast performance based on LS and QR. The seed is obtained from *Swedish female* mortality data from 1907 to 2006. Forecast errors of RMSE, MAPE, and bias for $e_0(t)$ for 60-year horizons are presented for both Case 1 (left column panel) and Case 2 (right column panel).

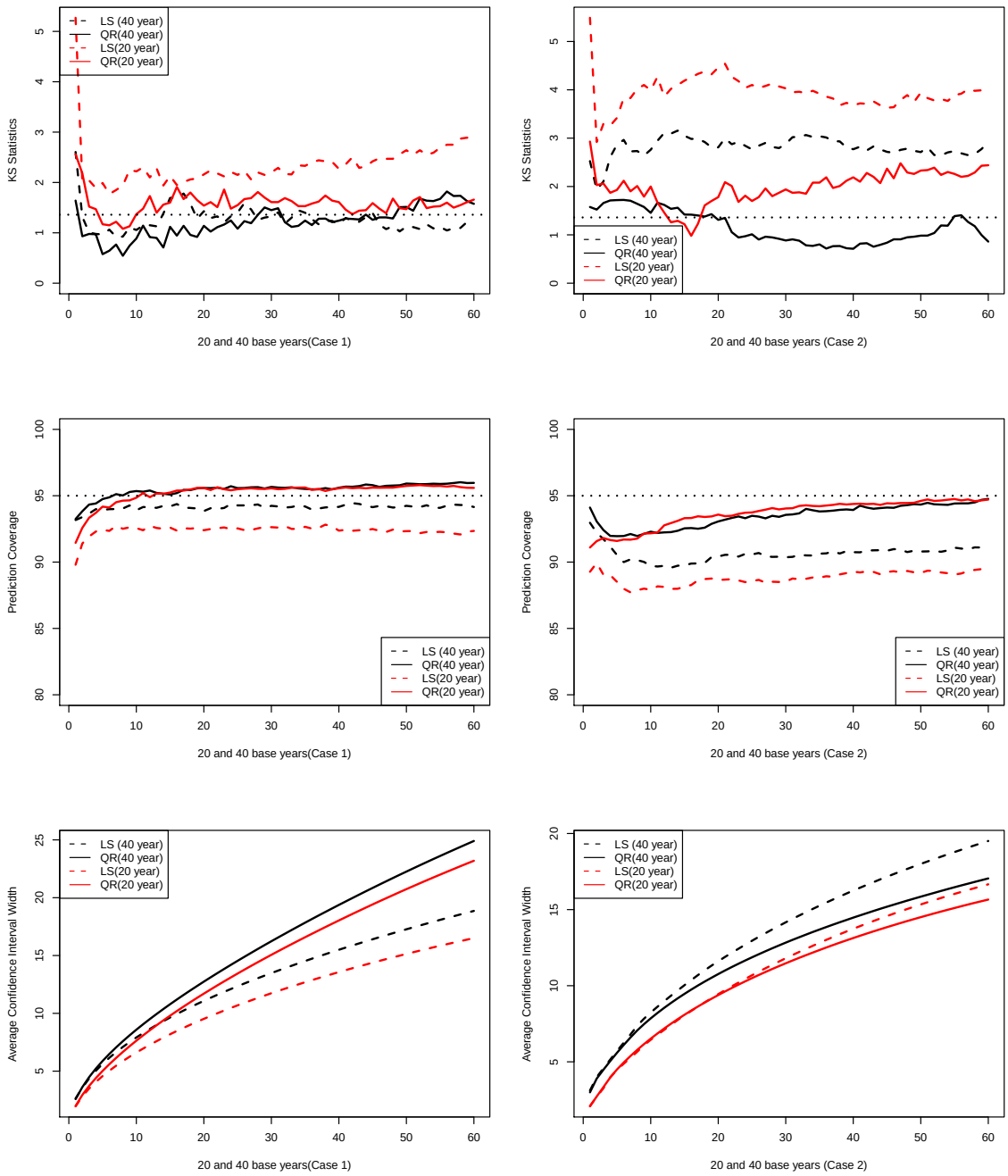

Figure 5: Simulation study on forecast performance based on LS and QR. The seed is obtained from *Swedish female* mortality data from 1907 to 2006. Kolmogorov-Smirnov statistic, coverage and average confidence interval width for $e_0(t)$ for 60-year horizons are presented for both Case 1 (left column panel) and Case 2 (right column panel). The critical value 1.36 from Kolmogorov-Smirnov test and the nominal coverage 95 percent are also marked.

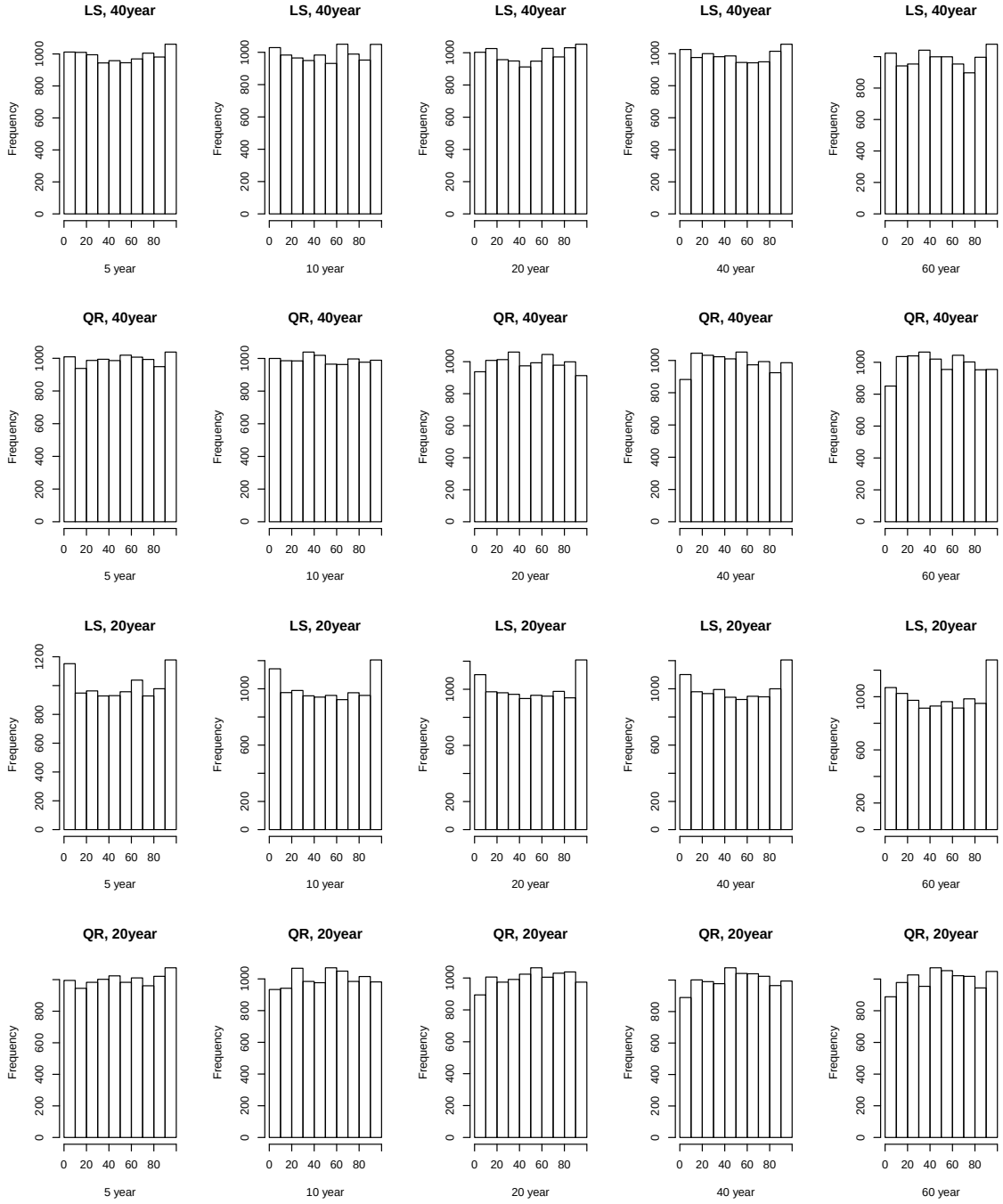

Figure 6: Simulation study on forecast performance based on LS and QR. The seed is obtained from *Swedish female* mortality data from 1907 to 2006. Percentile histograms for forecasts $e_0(t)$ for five, 10-, 20-, 40- and 60-year horizons are presented for *Case 1* for LS (40-year base), QR (40-year base), LS (20-year base), and QR (20-year base), respectively.

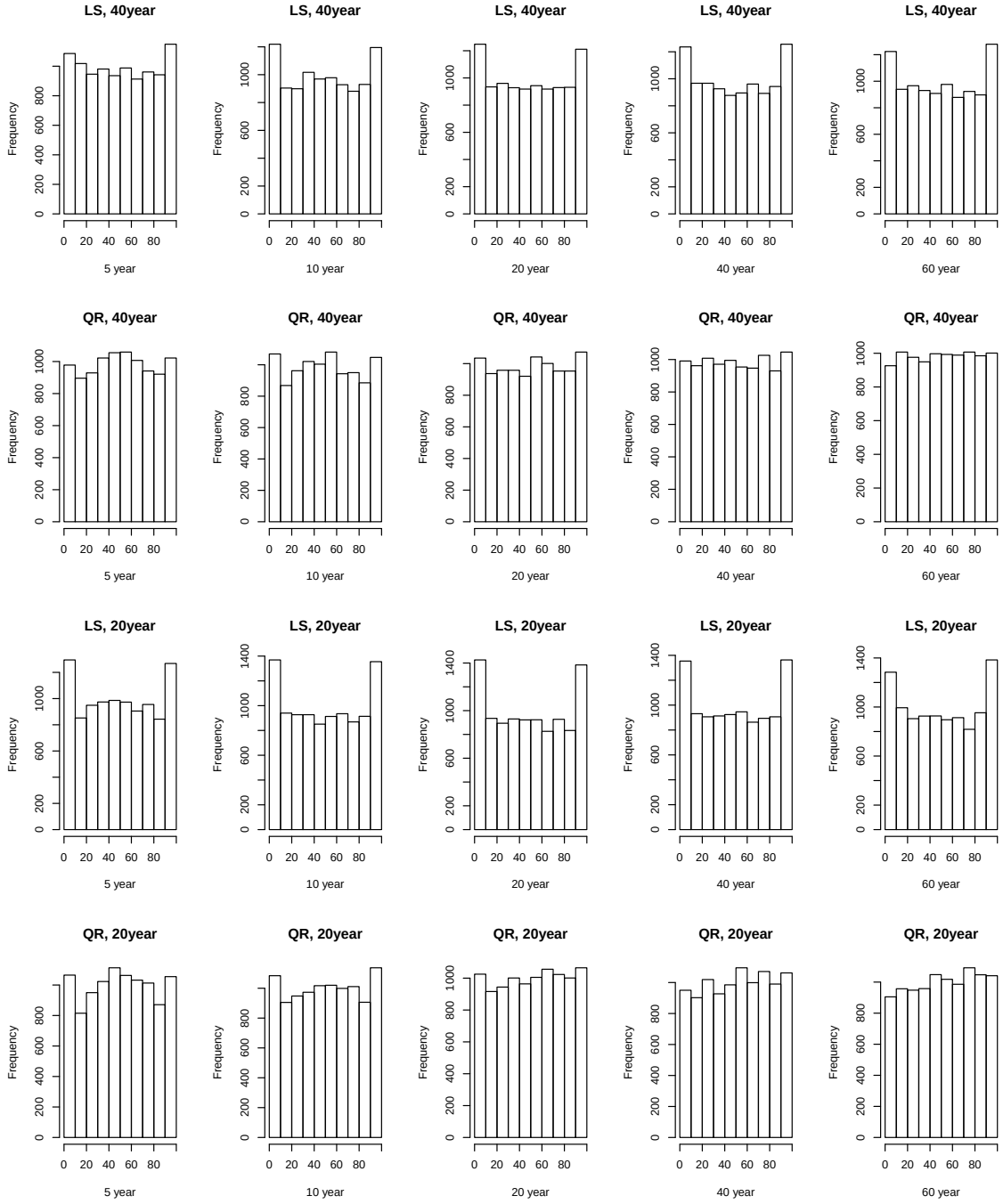

Figure 7: Simulation study on forecast performance based on LS and QR. The seed is obtained from *Swedish female* mortality data from 1907 to 2006. Percentile histograms for forecasts $e_0(t)$ for five-, 10-, 20-, 40- and 60-year horizons are presented for *Case 2* for LS (40-year base), QR (40-year base), LS (20-year base), and QR (20-year base), respectively.

5 Real data analysis

Based on the same idea of needing to construct a sample of forecasts to make a more formal statistical test, we implement the rolling window procedure in the analysis of real data. For example, consider the Canadian mortality data from 1921 through 2006. Let the base period length be fixed at 40 years. The forecasts are then based on data from $[1921 + i, 1960 + i]$, where $i = 0, \cdots, 44$. In total, we have made 45 consecutive forecasts with the 40-year base period. Each time, the jump off year moves forward one year. As a result, we have 45 one-year horizon forecasts, 44 two-year horizon forecasts, and so on. We examine the forecasts up to a 40-year horizon, keeping the sample size of at least six in all comparison.

Figures 8 to 11 display the results based on Canadian mortality data from 1921 to 2006, both female and male, respectively. As before, we present the forecast errors comprising RMSE, MAPE and bias, and density forecast performance criteria including Kolmogorov-Smirnov statistics, coverage and average confidence interval width. We also put together the results from implementing LS and QR so we can compare the impact of two methods. Again, solid lines are from QR, and dashed lines from LS. We have carried out similar analysis on Swedish and U.K. mortality data. Additional results are provided in Appendix B. There are some interesting features to note.

- Male forecasts are poorer in general than female's for the data we have examined in this study. This is obvious if we compare figures 8 and 9, and then figures 10 and 11 for each criterion we have computed.

- While there is no clear advantage of using a shorter base period for the females, it seems obvious that a shorter base period prediction generates better results for the males (see black lines for the 30-year based results v.s. red lines for the 40-year based results).

We think all of the differences are rooted in the possible structural change in male mortality improvement. See Figure 12 for the fitted Lee-Carter model components over the examined history period for female and male. Comparing the time-varying index k_t for female and male, we find that female k_t shows a rather consistently linear declining pattern, but the k_t pattern for the male is curved, implying trend changes. When structural changes are present, a shorter base

period can follow the underlying pattern faster, therefore giving better forecast performance. The phenomenon of "shorter base period, better prediction performance" is actually evidence for structural change, in our opinion.

Furthermore, the presence of structural changes implies the failure of the Lee-Carter model. If structural changes happen, the Lee-Carter model is not able to provide adequate prediction any more. QR, as a robust method to LS under the framework of the Lee-Carter model, doesn't apply either. In other words, QR can not remedy the abnormal situation due to structural changes. As a result, we don't see any significant improvement from QR in the real data analysis of male mortality prediction.

This should be interpreted differently for Canadian female mortality data, where the improvement from QR prediction is also very limited. The reason for female mortality is probably that there is no (or little) structural change in the time-varying index k_t, nor irregular large shocks in k_t's fluctuation during the studied period. Therefore, the improvement from QR is very minor, similar to Case 1 in the simulation study. We have similar observations for Swedish and U.K. mortality prediction—both male and female.

For Swedish female and male mortality prediction (see figures 26 and 27), it is interesting to note that the confidence intervals from QR are much narrower than those from LS, while the Kolmogorov-Smirnov test and coverage from QR still perform similarly to those from LS. We think this difference is due to the effect of the 1918 Spanish flu pandemic in Swedish mortality data, and QR presents robustness in dealing with large shocks.

We also compare forecast performance between the simulation study and the real data analysis. Table 1 puts together the various forecast assessment criteria for forecasts $e_0(t)$ at 20- and 40-year horizons. In the simulation, we use Canadian female and male mortality patterns from 1921 to 2006 as seeds to generate different scenarios (Case 1 with the LS method used). Therefore, the results reflect the overall average for each criterion. The real data results are based on the rolling procedure of a 40-year base period. We see that the errors resulting from female forecasts are at similar magnitudes (except that bias becomes 10 times of the simulated one at 40-year horizon). However, the errors in male forecasts are at a magnitude that is from triple to nine times of their counterparts from the simulation, and coverage drops to 11.11 percent at 20-year horizon, and 0 percent at 40-year horizon with real data. This seems to

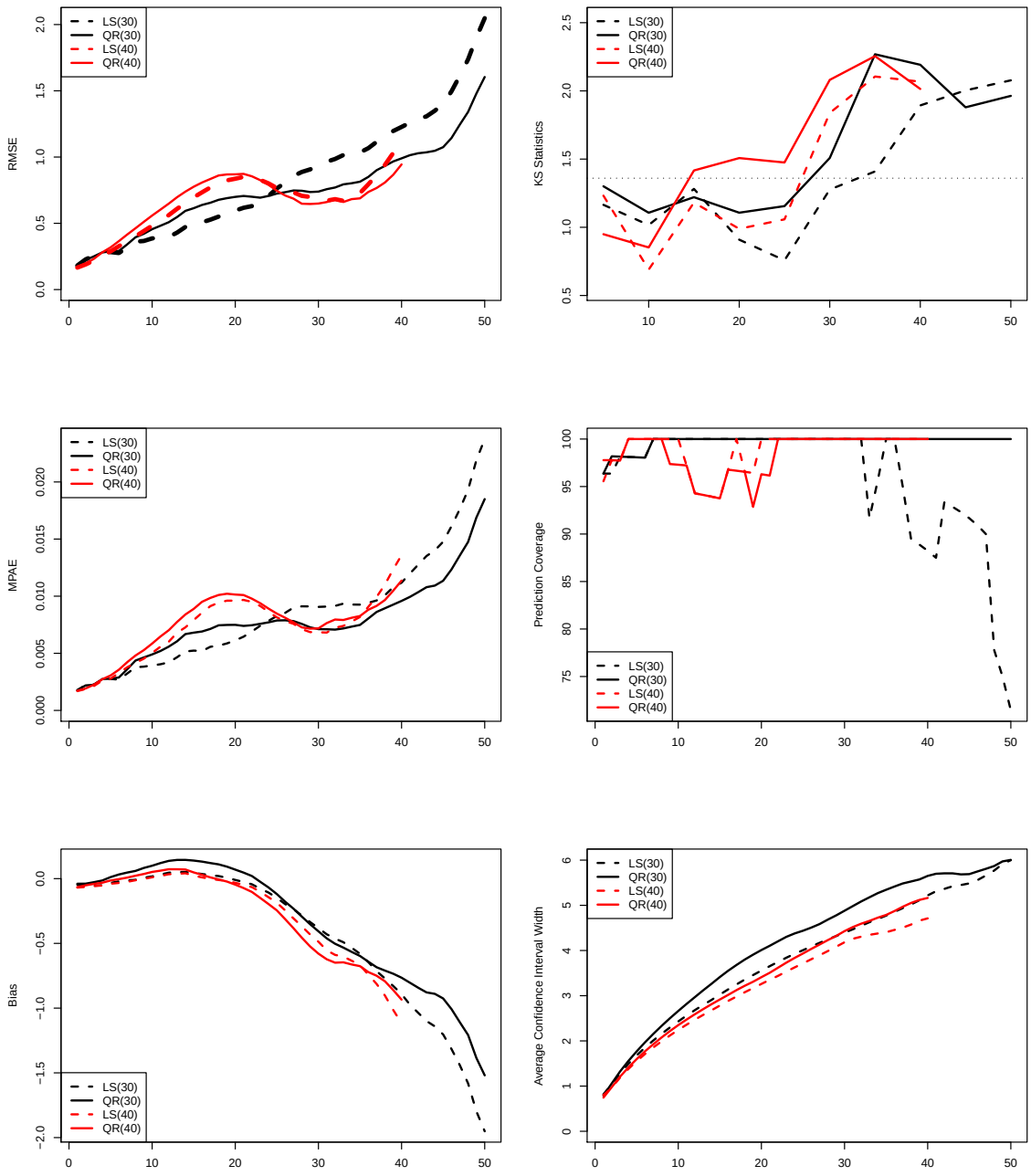

Figure 8: Real data analysis on Canadian female mortality data from 1921 to 2006 based on LS and QR using 30- or 40-year base period. Rolling window procedure is applied and the first jump-off year is 1960. Presented are forecast criteria: RMSE, MAPE, bias, Kolmogorov-Smirnov statistic, coverage and average confidence interval width for $e_0(t)$.

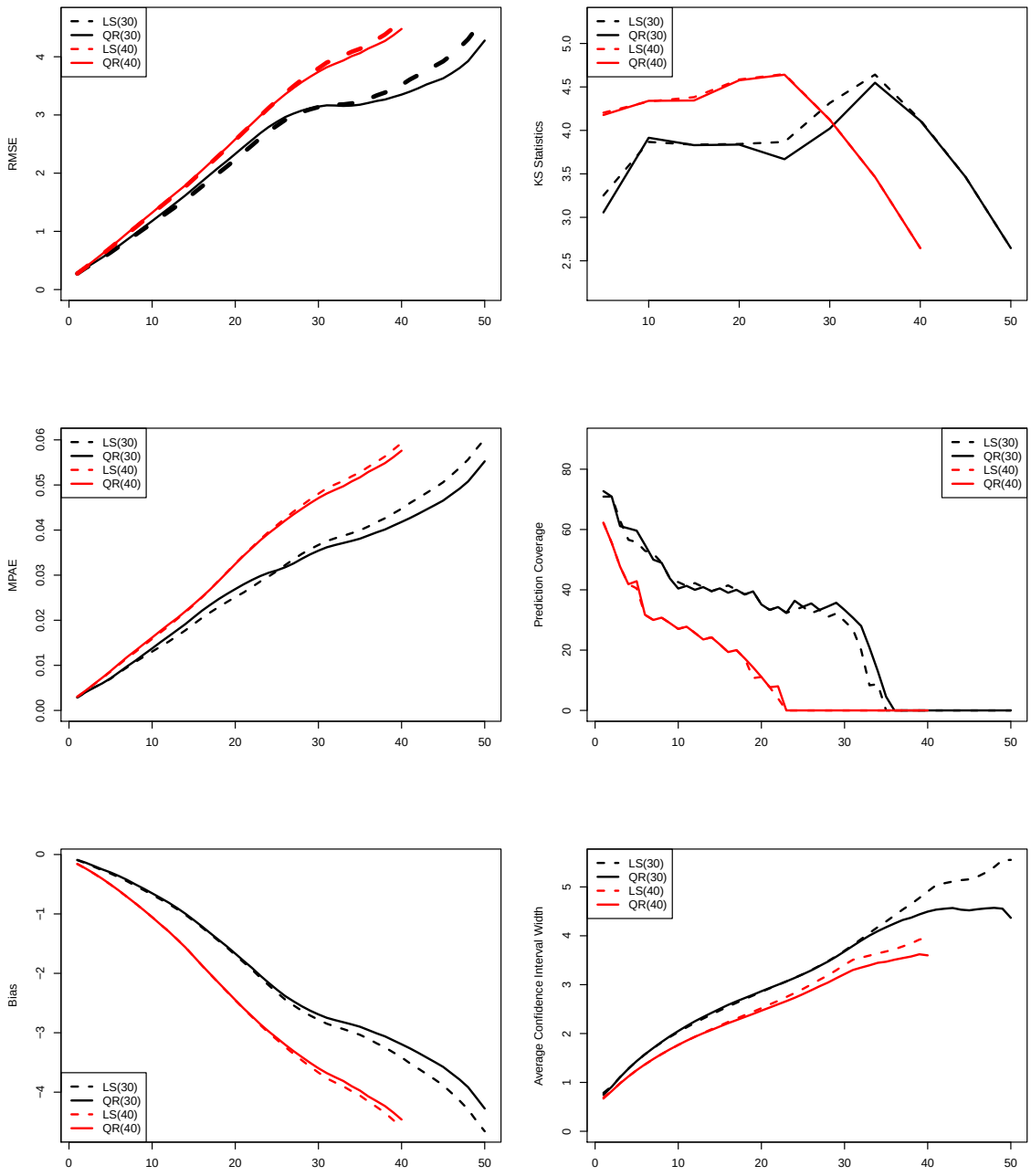

Figure 9: Real data analysis on Canadian male mortality data from 1921 to 2006 based on LS and QR using 30- or 40-year base period. Rolling window procedure is applied and the first jump-off year is 1960. Presented are forecast criteria: RMSE, MAPE, bias, Kolmogorov-Smirnov statistic, coverage and average confidence interval width for $e_0(t)$.

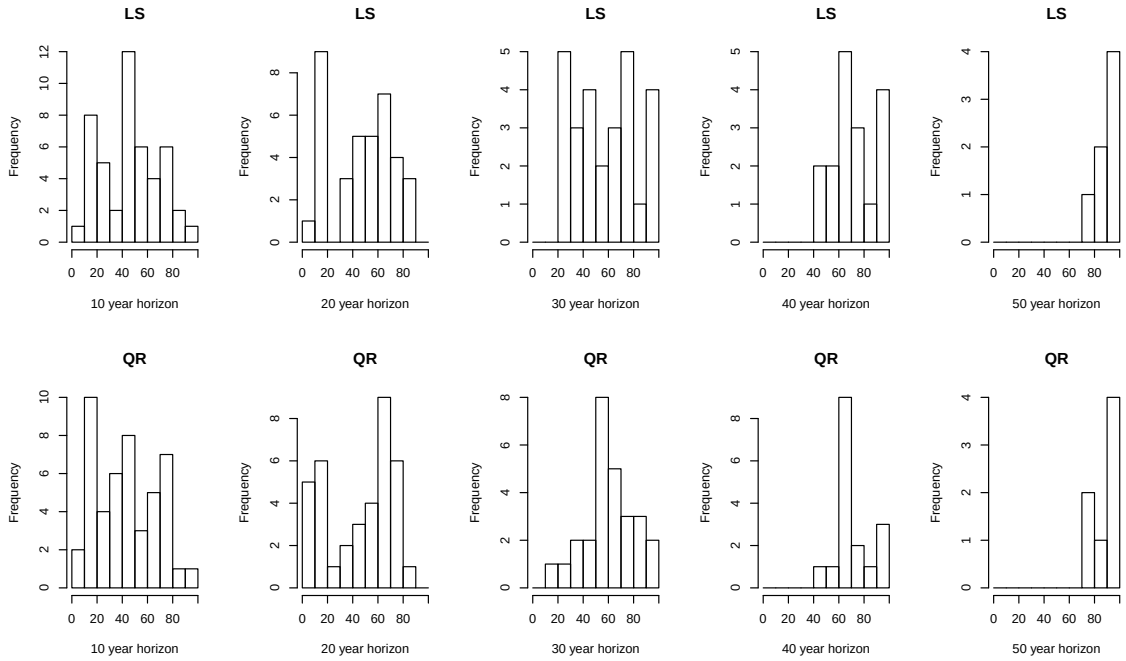

Figure 10: Real data analysis on Canadian female mortality data from 1921 to 2006. Rolling window procedure is applied, the base period is 30 years and the first jump off year is 1960. Presented is the percentile histograms from LS and QR

suggest that the documented underprediction in the Lee-Carter model application is the result of model mis-specification.

Due to moving base and uneven size by forecast horizon, statistical evaluation of prediction performance should be interpreted with caution. This is particularly true for long-horizon predictions because as few as six samples are used to compute SMSE, bias, etc.

6 Conclusion

In this paper, we set up a backtesting framework to assess long-term prediction performance of the Lee-Carter model. We emphasize the importance of examining two aspects—the accuracy of point forecasts and the adequacy of distribution forecasts equally. Various forecast performance criteria are used. In addition to the commonly used error measurements RMSE, MAPE, bias and so on, we also carry out the goodness-of-fit test based on Kolmogorov-Smirnov statistic.

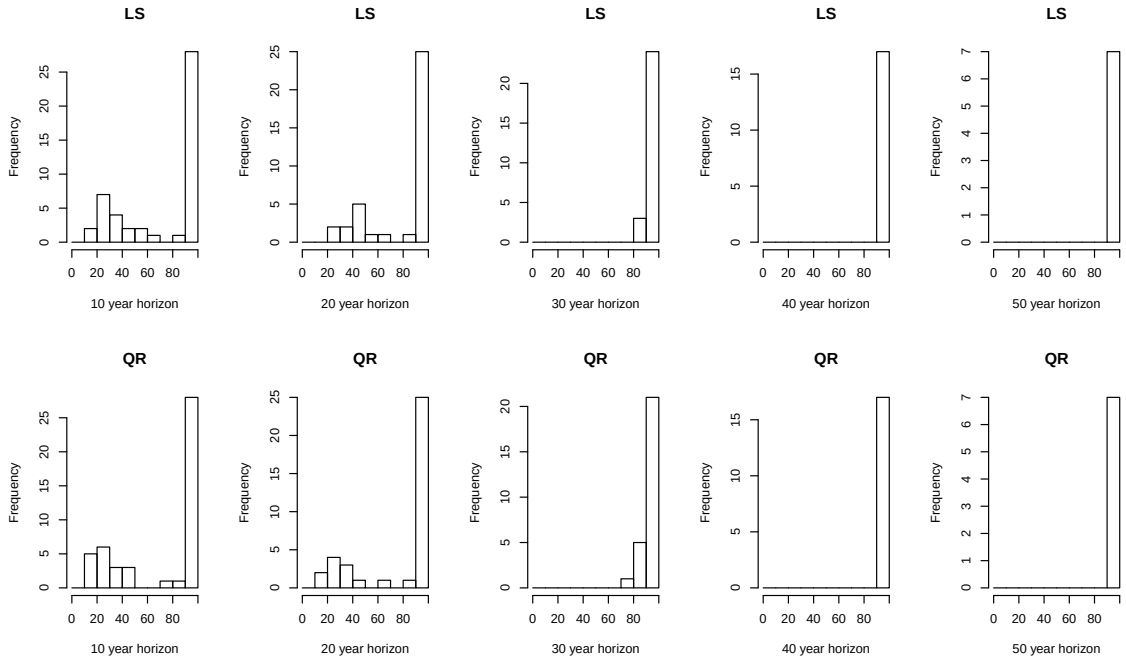

Figure 11: Real data analysis on Canadian male mortality data from 1921 to 2006. Rolling window procedure is applied, the base period is 30 years and the first jump-off year is 1960. Presented is the percentile histograms from LS and QR

This is to check if the realized outcomes agree with the predicted distribution.

The time-varying index k_t is critical for Lee-Carter model prediction. In the evolution of human mortality, large variations occur once in a while. To improve the robustness of the model parameter estimation under different conditions, we propose to use quantile regression to the model of k_t. We have shown that QR improves prediction performance when data contain irregular shocks through simulation study in Section 4.3. Even though QR may not perform as well as we expected for real data, we believe QR is a safe-guard method against any hidden abnormal situations in real data. LS and QR should complement each other—the two methods should be used together and any large discrepancy between the results of applying these two methods should be investigated further.

We closely examine the issue of taking into account drift uncertainty in deriving the forecast distribution. Using our designed simulation study, it is shown that the drift term in k_t plays a very important role in generating plausible prediction intervals, particularly when we are

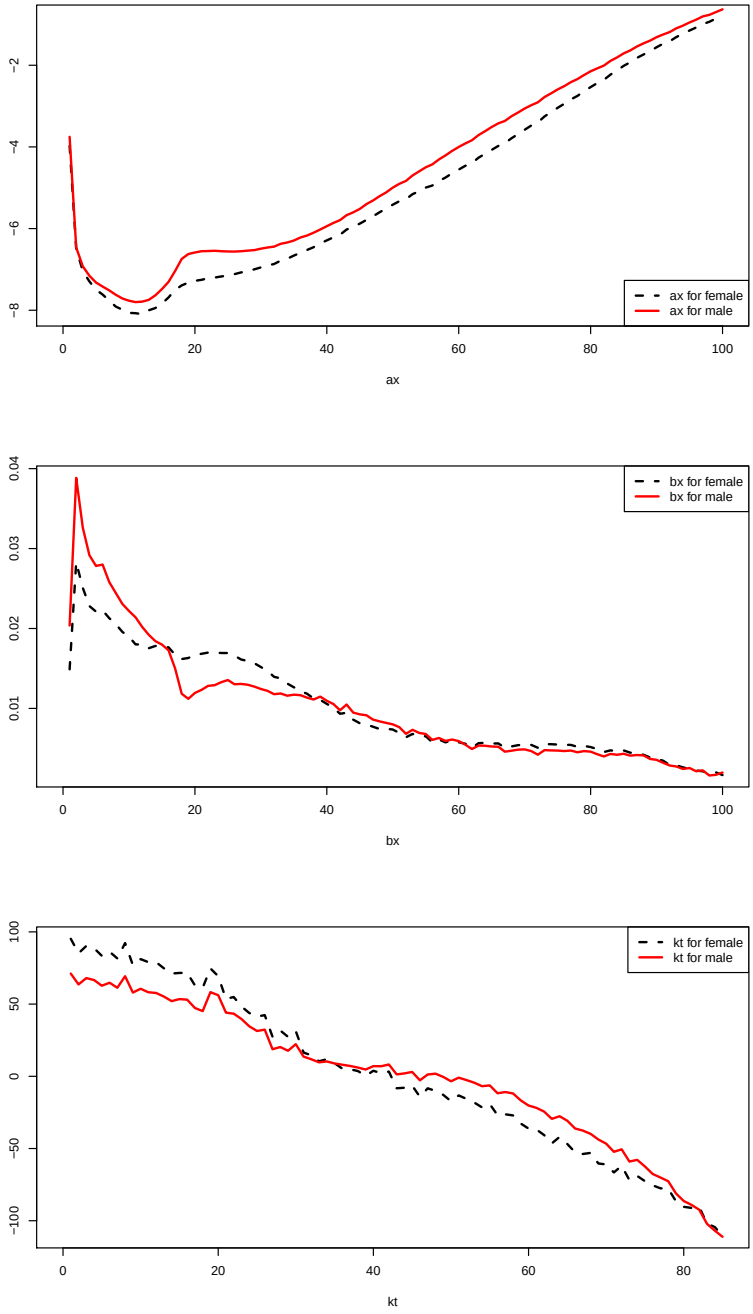

Figure 12: The fitted Lee-Carter model estimates based on Canadian female and male mortality data from 1921 to 2006.

interested in long-term predictions. Our proposed bootstrap procedure is able to capture the uncertainty in the drift term, distribution-free. Our results reveal ignoring drift uncertainty in

Table 1: Forecast criteria at 20- and 40-year horizons in simulation and real data analysis

	20-year horizon comparison			
	Can. female (sim)	Canadian female	Can. male (sim)	Canadian male
RMSE	1.166	0.8365	0.9885	2.569
MAPE	0.0117	0.0096	0.0104	0.0326
Bias	−0.0605	−0.0325	−0.0399	−2.4539
KS.test	0.71	0.9888	1.03	4.5857
Coverage	94.42%	100%	93.93%	11.11%
	40-year horizon comparison			
RMSE	1.4826	1.1307	1.2685	4.6083
MAPE	0.0142	0.0136	0.0129	0.0593
Bias	−0.1022	−1.1173	−0.0586	−4.5919
KS.test	0.8	2.0674	0.91	2.6458
Coverage	94.57%	100%	94.11%	0%

prediction may lead to severe underestimation to the future uncertainty in mortality.

Another innovation of this study is to make use of specially designed simulation schemes to test the proposed methodologies. Our simulation study allowed us to separate the model effectiveness issue from the model identification issue. We have shown that the forecast performance of the Lee-Carter model can be improved significantly with the QR method, particularly for the distribution prediction. In addition, the simulation reveals that there are systematic biases in predicting life expectancies using the Lee-Carter model. This bias need to be addressed in order to have correct predictions. The causes of the bias will be explored in future work.

Finally, with real data analysis, female and male mortality data normally exhibit very different predictability. This is true for all the data sets we have examined: from Sweden, the United Kingdom and Canada. Compared to what we have seen in the simulated data, there are potential trends or structural changes in male mortality. This could certainly be a reason projections from a shorter base period outperform those with longer base periods. This also means that, in reality, a simple Lee-Carter model covering the whole range of observation may

mis-specify the process that drives mortality dynamics. The results with real data analysis raise caution to users of the Lee-Carter model. In practice, users tend to use the data with a period as short as 10 years to calibrate the Lee-Carter model; the long-term prediction based on this method could be questionable.

Acknowledgments

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada. The authors thank John Braun for his help during the writing of the paper, and two anonymous referees from the Society of Actuaries Living to 100 Symposium Committee for valuable comments and suggestions. We also acknowledge Sirlei A. Cavassin for her research assistance.

The majority of simulations was carried out on SHARCNET (http://www.sharcnet.ca). The use of the SHARCNET facility is greatly appreciated.

References

Bell, W. R. 1997. "Comparing and Assessing Time Series Methods for Forecasting Age-Specific Fertility and Mortality Rates." *Journal of Official Statistics* 13: 279–303.

Bilingsley, P. 1999. "Convergence of Probability Measures." 2nd Edition. Wiley, New York.

Booth, H., Maindonald, J., and Smith, L. 2002. "Applying Lee-Carter Under Conditions of Variable Mortality Decline." *Population Studies* 56: 325–336.

Brouhns, N., Denuit, M., and Vermunt, J. 2002. "A Poisson Log-Bilinear Regression Approach to the Construction of Projected Lifetables." *Insurance: Mathematics and Economics* 31: 373–393.

Cairns, A. J. G., Blake, D., and Dowd, K. 2006. "Pricing Death: Frameworks for the Valuation and Securitization of Mortality Risk" *ASTIN Bulletin* 36: 79–120.

Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., and Epstein, D. 2009. "A Quantitative Comparison of Stochastic Mortality Models Using Data from England and Wales and the United States." *North American Actuarial Journal* 13: 1–35.

Carter, L. R. 1996. "Forecasting U.S. Mortality: A Comparison Of Box-Jenkins ARIMA and Structural Time Series Models." *The Sociological Quarterly* 37: 127–144.

Continuous Mortality Investigation 2005. "Projecting Future Mortality: Towards a Proposal For a Stochastic Methodology" CMI Working Paper 15.

Denuit, M. 2008. "Comonotonic Approximations to Quantiles of Life Annuity Conditional Expected Present Values" *Insurance: Mathematics and Economics* 42: 831–838.

Dowd, K., Cairns, A. J., Blake, D., Coughlan, G. D., Epstein, D., and Khalaf-Allah, M. 2008. "Backtesting Stochastic Mortality Models: An Ex-Post Evaluation of Multi-Period-Ahead Density Forecasts" DISCUSSION PAPER PI-0803, Pensions Institute.

Koenker, R. and Bassett, G. 1978. "Regression Quantiles." *Econometrica* 46: 33–50.

Koenker, R. and Hallock, K. F. 2001. "Quantile Regression." *Journal of Economic Perspectives* 15: 143–156.

Koul, H. L. 2002. "Weighted Empirical Processes in Dynamic Nonlinear Models." *Lecture Notes in Statistics* 166 2nd Edition. Springer-Verlag.

Lee, R. D. and Carter, L. R. 1992. "Modeling and Forecasting U.S. Mortality." *Journal of the American Stotistical Association* 87: 659–675.

Lee, R. D. and Miller, T. 2001. "Evaluating the Performance of the Lee-Carter Method for Forecasting Mortality" *Demography* 38: 537–549.

Liu, X. and Braun, W. J. 2010. "Investigating Mortality Uncertainty Using the Block Bootstrap." *Journal of Probability and Statistics* 2010: 15 pages.

Pitacco, E. 2004. "Survival Models in a Dynamic Context: A Survey." *Insurance: Mathematics and Economics* 35: 279–298.

Renshaw, A. and Haberman, S. 2003). "Lee-Carter Mortality Forecasting with Age-Specific Enhancement" *Insurance: Mathematics and Economics* 33: 255–272.

— 2006. "A Cohort-Based Extension to the Lee-Carter Model for Mortality Reduction Factors." *Insurance: Mathematics and Economics* 38: 556–570.

Yu, H. 2002. "Rmpi: Parallel Statistical Computing in R. " *R News* 2: 10–14.

A Kolmogorov-Smirnov test

In this section, we derive the asymptotic distribution of *ks.statistic* used in (8). First we assume that $X_1, \ldots, X_n$ are n variables at a certain time of future. Each X_k follows a continuous distribution function $F_k(x)$ which is unknown to us. But we can simulate m iid random variables $X_{k1}, \ldots, X_{km}$ that follow the same distribution as X_k. Define the empirical distribution function of $\{X_{k1}, \ldots, X_{km}\}$ as

$$F_{k,m}(x) = \frac{1}{m} \sum_{i=1}^{m} I(X_{ki} \leq x).$$

We can then use $F_{k,m}(x)$ to define the sample percentile of X_k given $\{X_{k1}, \ldots, X_{km}\}$ as $p_k = F_{k,m}(X_k)$. Notice that $0 \leq p_k \leq 1$ and $\sum_{i=1}^{m} I(X_{ki} \leq x)$ follows binomial distribution $Bin(m, F_k(x))$ for each fixed x. Thus the distribution function of p_k is

$$\begin{aligned}
U_{k,m}(s) &= P\{p_k \leq s\} \\
&= E[E[I(F_k(X_k) \leq s)|X_k]] \\
&= E[E[I\Big(\sum_{i=1}^{m} I(X_{ki} \leq X_k) \leq ms\Big)|X_k]] \\
&= E\sum_{i=0}^{[ms]} \binom{m}{i} F_k^i(X_k)(1 - F_k(X_k))^{m-i} \\
&= \sum_{i=0}^{[ms]} \binom{m}{i} Beta(i+1, m-i+1) \\
&= \frac{[ms]+1}{m+1}, \ 0 \leq s \leq 1,
\end{aligned}$$

where we use the fact that $F_k(X_k)$ follows uniform[0,1] distribution so the expectation links to Beta function. One obvious conclusion is that $U_{k,m}(s)$ is independent of $F_k(x)$. In addition, it

is easy to check

$$\sup_{0\le s\le 1} |U_{k,m}(s) - s| \le 1/(m+1). \tag{A.1}$$

So

$$\lim_{m\to\infty} \sup_{0\le s\le 1} |U_{k,m}(s) - s| = 0. \tag{A.2}$$

This leads to the result in the following Proposition.

Proposition A.1 *p_k follows uniform[0,1] distribution asymptotically.*

Since $X_{k1}, \ldots, X_{km}$ are simulated from the specific model and m can be as large as possible, we assume that m is a function of n with constrain $m \ge n$.

Define the empirical distribution function based on $p_k, k = 1, \ldots, n$ as

$$\hat{F}_n(s) = \frac{1}{n}\sum_{k=1}^{n} I(p_k \le s),\ 0 \le s \le 1$$

and corresponding empirical process as

$$B_n(s) = \sqrt{n}(\hat{F}_n(s) - s),\ 0 \le s \le 1.$$

Definition A.2 *A process $\{B(s), 0 \le s \le 1\}$ is called a Brownian bridge if it is a mean zero Gaussian process with covariance $EB(s)B(t) = \min\{s,t\} - st,\ 0 \le s,t \le 1$.*

The following result involves convergence of distribution in infinite dimensional space. Many technique terminologies such as Brownian bridge and Skorokhod topology can be found in Billingsley (1999).

Proposition A.3 *Assume that $\{X_1, \ldots, X_n\}$ are independent of each other for $k = 1, 2, \ldots, n$. Then, as $n \to \infty$, $\{B_n(s), 0 \le s \le 1\}$ converges weakly in the Skorokhod space $\mathcal{D}[0,\ 1]$ to $\{B(s), 0 \le s \le 1\}$.*

Proof: Define

$$W_n(s) = \sum_{k=1}^{n} n^{-1/2}\big(I(p_k \le s) - U_{k,m}(s)\big),\ 0 \le s \le 1.$$

Then $B_n(s)$ can be expressed by two parts:

$$B_n(s) = W_n(s) + \sum_{k=1}^{n} n^{-1/2}(U_{k,m}(s) - s),\ 0 \le s \le 1.$$

Since by (A.1)

$$\sup_{0\le s\le 1}\left|\sum_{k=1}^{n} n^{-1/2}(U_{k,m}(s)-s)\right| \le \sqrt{n}/(m+1) \to 0 \text{ as } n\to\infty \,(\text{and } \; m\to\infty).$$

We just need to prove that $\{W_n(s)\}$ converges to $\{B(s)\}$ which follows by Theorem 2.2.1 in Koul (2002). This completes the proof of Proposition.

By using continuous mapping theorem (cf. Bilingsley (1999)), we obtain the asymptotic distribution of *ks.statistic*.

Corollary A.4 *We have*

$$\sup_{0\le s\le 1}|B_n(s)| \xrightarrow{\mathcal{D}} \sup_{0\le s\le 1}|B(s)|.$$

When $X_1,\ldots,X_n$ are not independent, the above result is no longer true. Given a set of suitable conditions such as stationarity of $X_1,\ldots,X_n$, $B_n(s)$ still converges to a Gaussian process but no longer to be a Brownian bridge.

B Additional results

We have implemented the evaluation procedures described in Sections 4 and 5 to data sets from different countries—Sweden, the United Kingdom, and Canada—and both genders. The overall observations are similar, with a few exceptions. Additional results not been given in the main text are attached here.

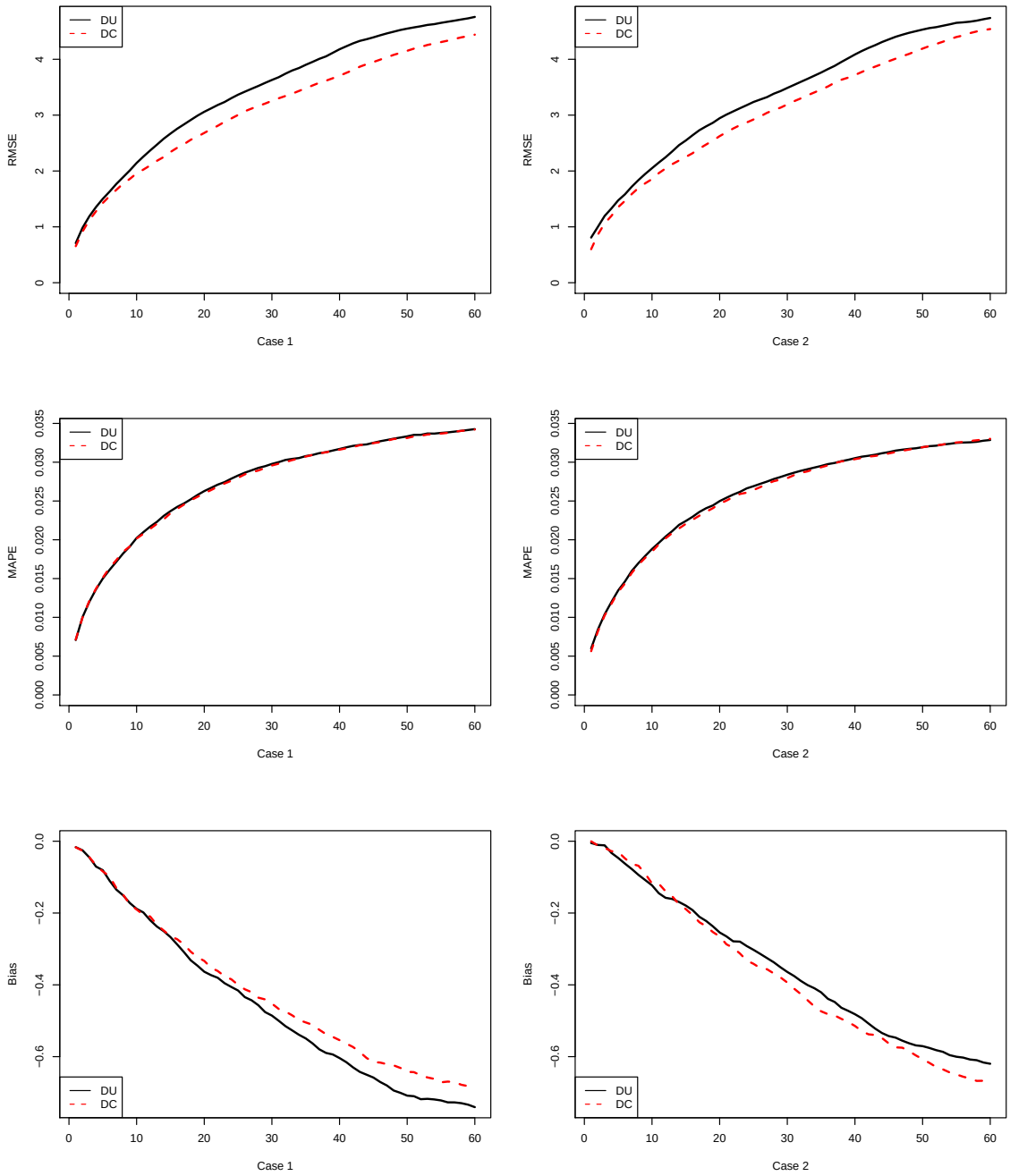

Figure 13: Simulation study over the effect of drift uncertainty on forecast performance. The seed is obtained from *Swedish male* mortality data from 1907 to 2006. Forecast errors of RMSE, MAPE, and bias for $e_0(t)$ for 60-year horizons are presented for both Case 1 (left column panel) and Case 2 (right column panel).

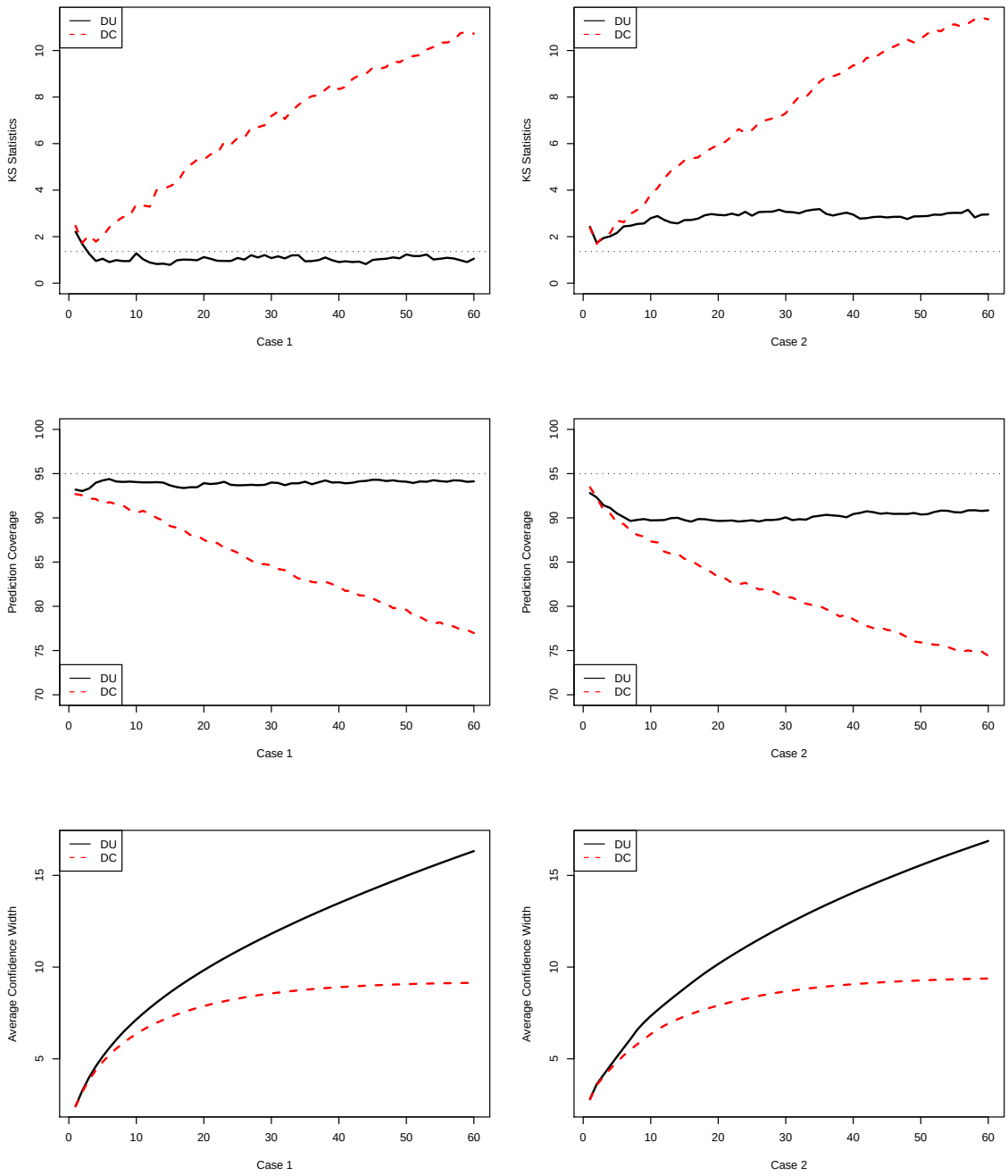

Figure 14: Simulation study over the effect of drift uncertainty on forecast performance. The seed is obtained from *Swedish male* mortality data from 1907 to 2006. Density forecast for $e_0(t)$ for 60-year horizons are assessed by Kolmogorov-Smirnov statistics, coverage and average confidence interval width for both Case 1 (left column panel) and Case 2 (right column panel). The critical value 1.36 from Kolmogorov-Smirnov test and the nominal coverage 95 percent are also marked.

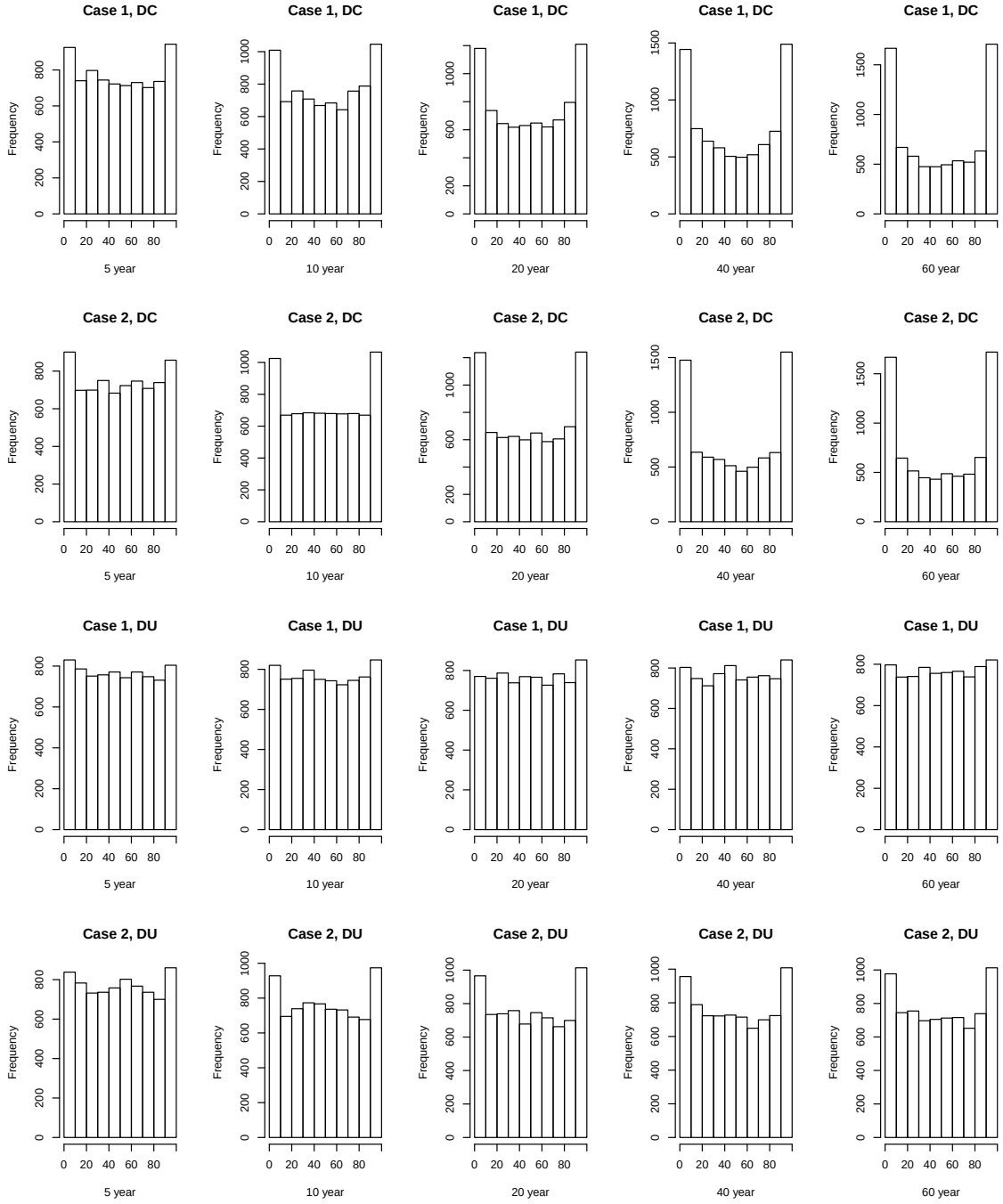

Figure 15: Simulation study over the effect of drift uncertainty on forecast performance. The seed is obtained from *Swedish male* mortality data from 1907 to 2006. Percentile histograms for forecasts $e_0(t)$ for five-, 10-, 20-, 40- and 60-year horizons are presented for Case 1 (DC), Case 2 (DC), Case 1 (DU) and Case 2 (DU), respectively.

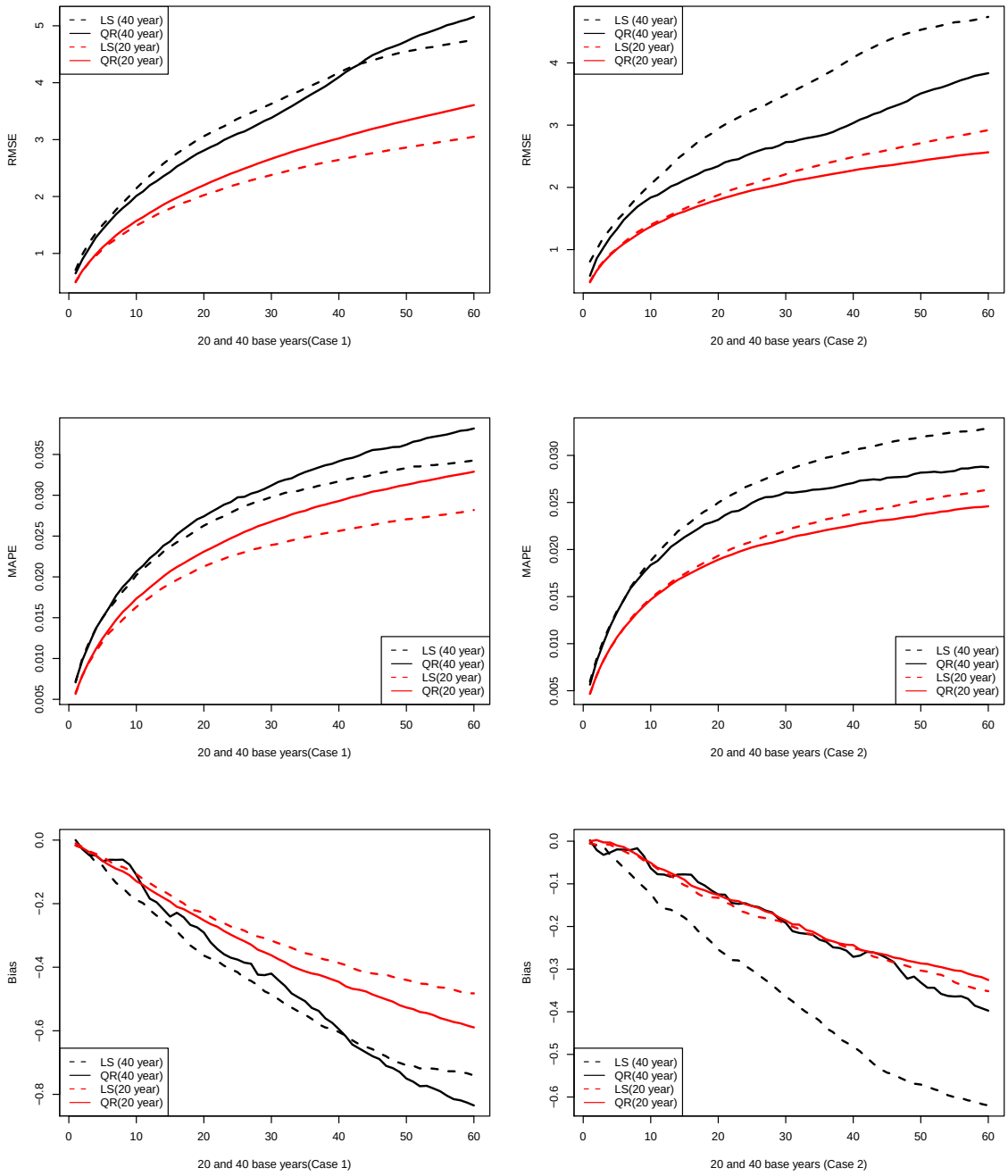

Figure 16: Simulation study on forecast performance based on LS and QR. The seed is obtained from *Swedish male* mortality data from 1907 to 2006. Forecast errors of RMSE, MAPE, and bias for $e_0(t)$ for 45-year horizons are presented for both Case 1 (left column panel) and Case 2 (right column panel).

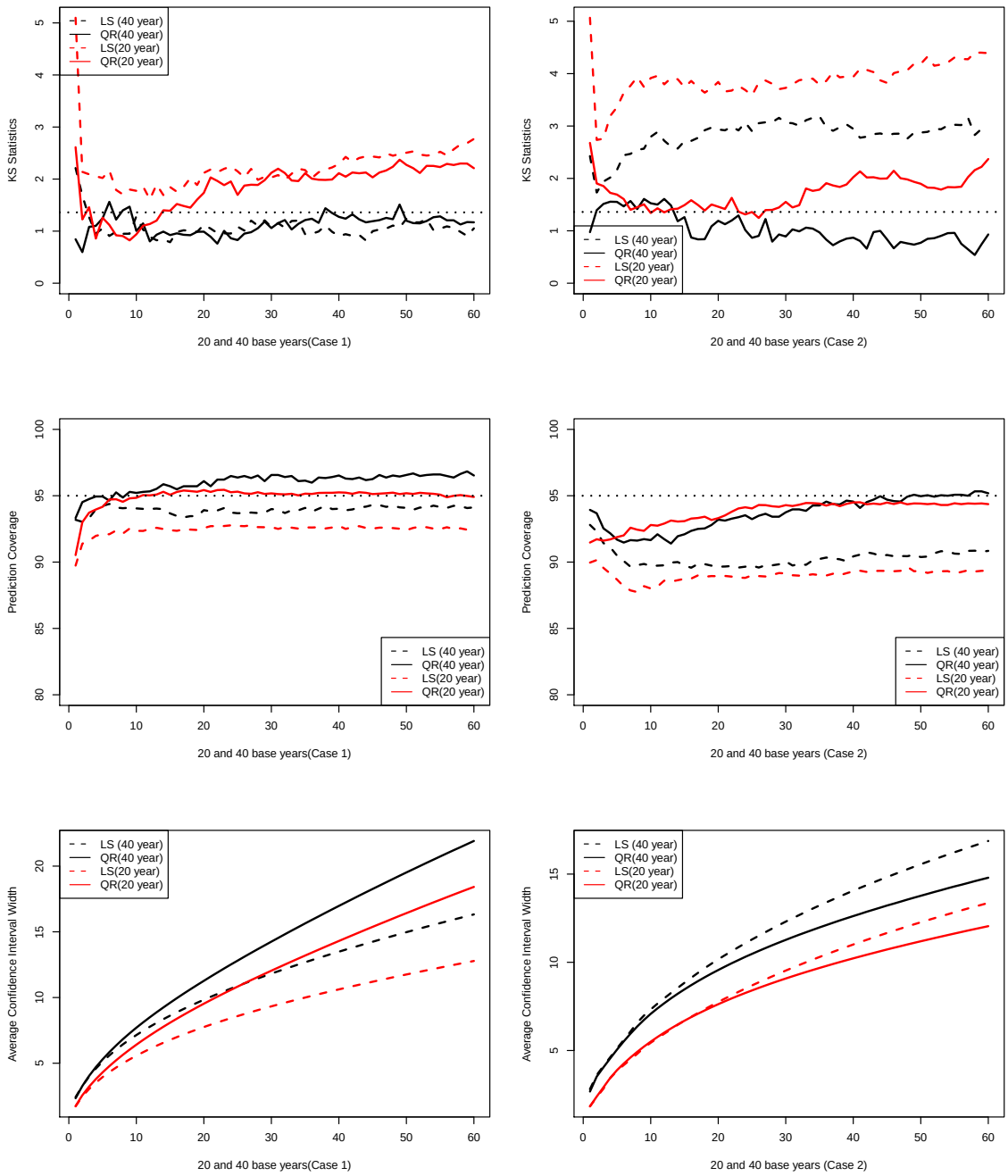

Figure 17: Simulation study on forecast performance based on LS and QR. The seed is obtained from *Swedish male* mortality data from 1907 to 2006. Kolmogorov-Smirnov statistic, coverage and average confidence interval width for $e_0(t)$ for 45-year horizons are presented for both Case 1 (left column panel) and Case 2 (right column panel). The critical value 1.36 from Kolmogorov-Smirnov test and the nominal coverage 95 percent are also marked.

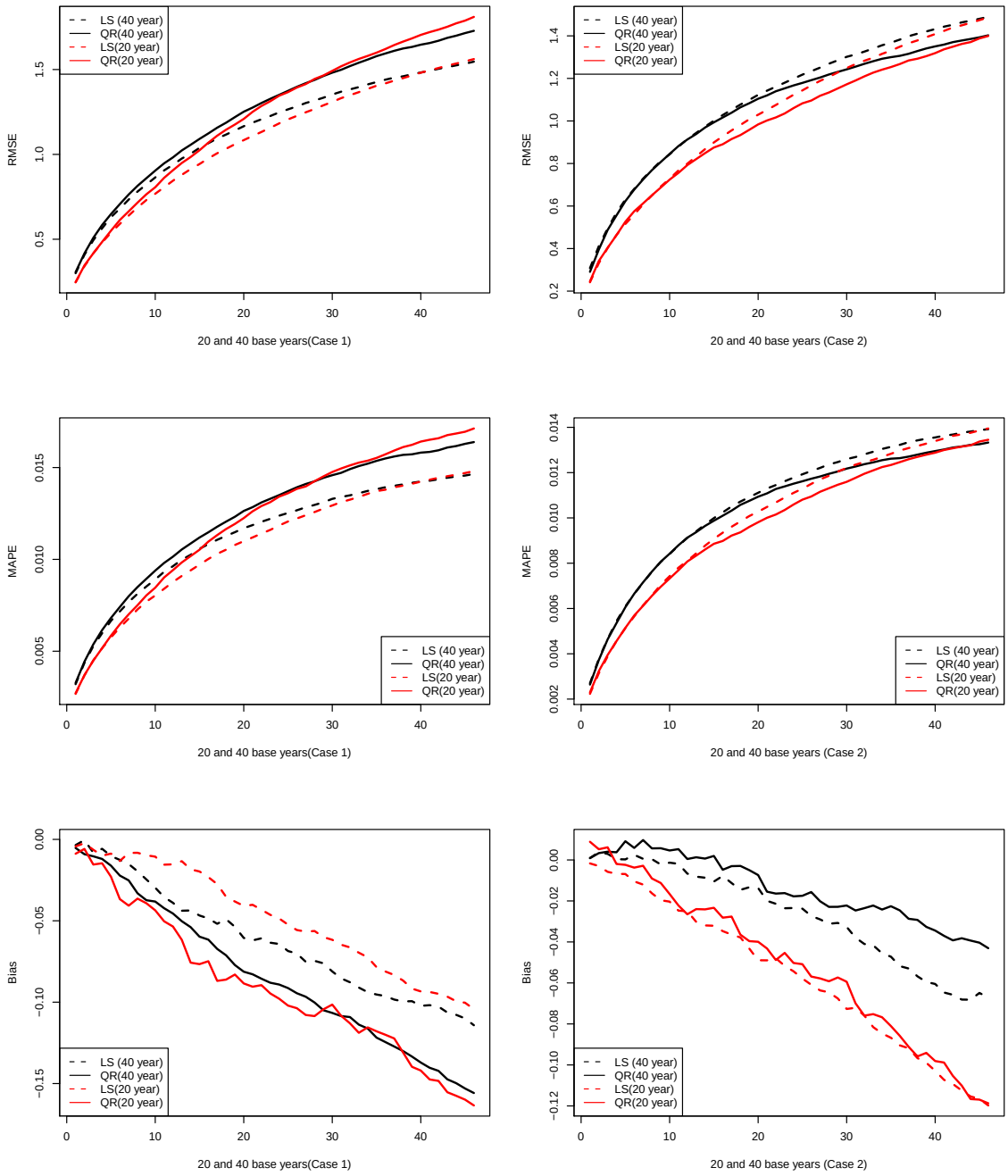

Figure 18: Simulation study on forecast performance based on LS and QR. The seed is obtained from *Canadian female* mortality data from 1921 to 2006. Forecast errors of RMSE, MAPE, and bias for $e_0(t)$ for 45-year horizons are presented for both Case 1 (left column panel) and Case 2 (right column panel).

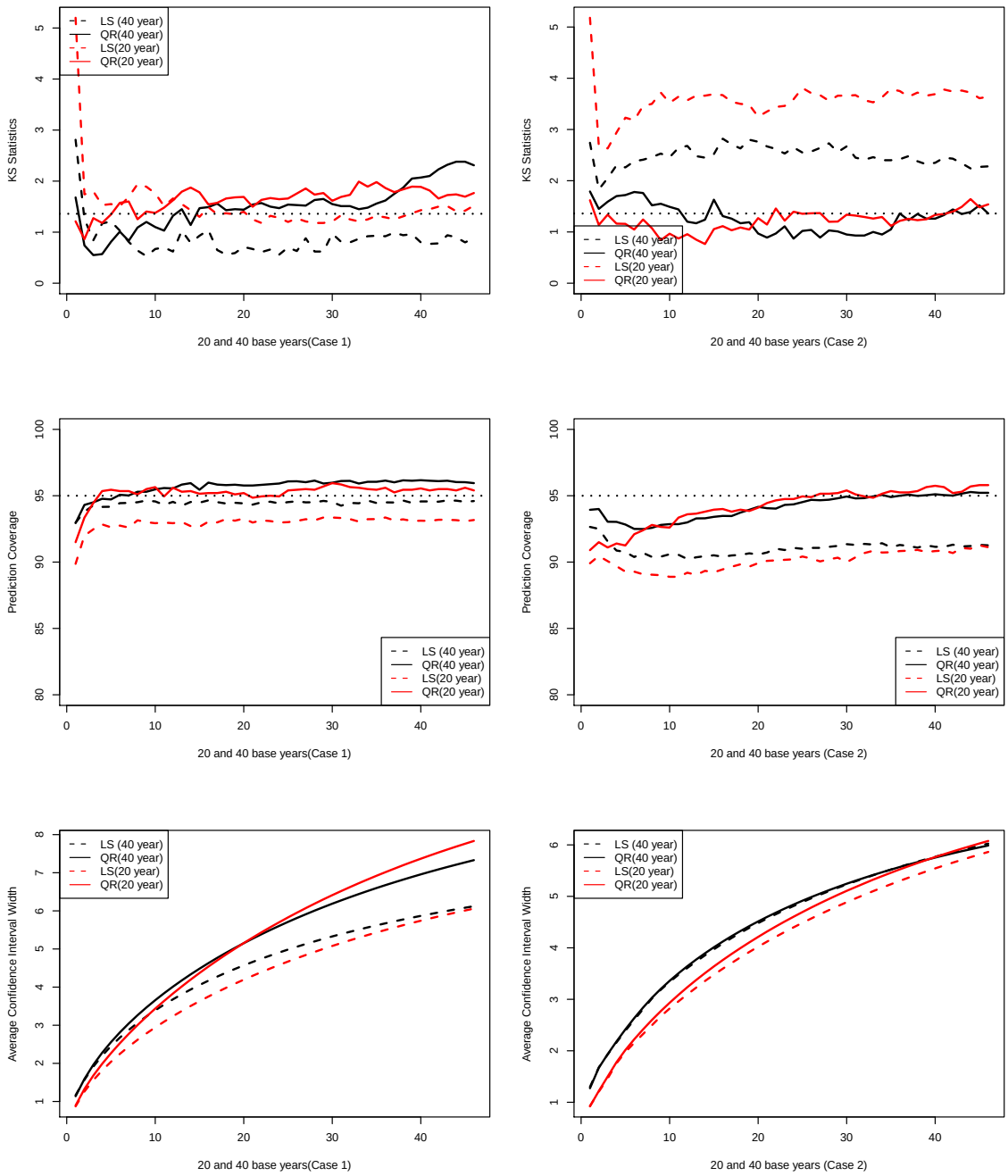

Figure 19: Simulation study on forecast performance based on LS and QR. The seed is obtained from *Canadian female* mortality data from 1921 to 2006. Kolmogorov-Smirnov statistic, coverage and average confidence interval width for $e_0(t)$ for 45-year horizons are presented for both Case 1 (left column panel) and Case 2 (right column panel). The critical value 1.36 from Kolmogorov-Smirnov test and the nominal coverage 95 percent are also marked.

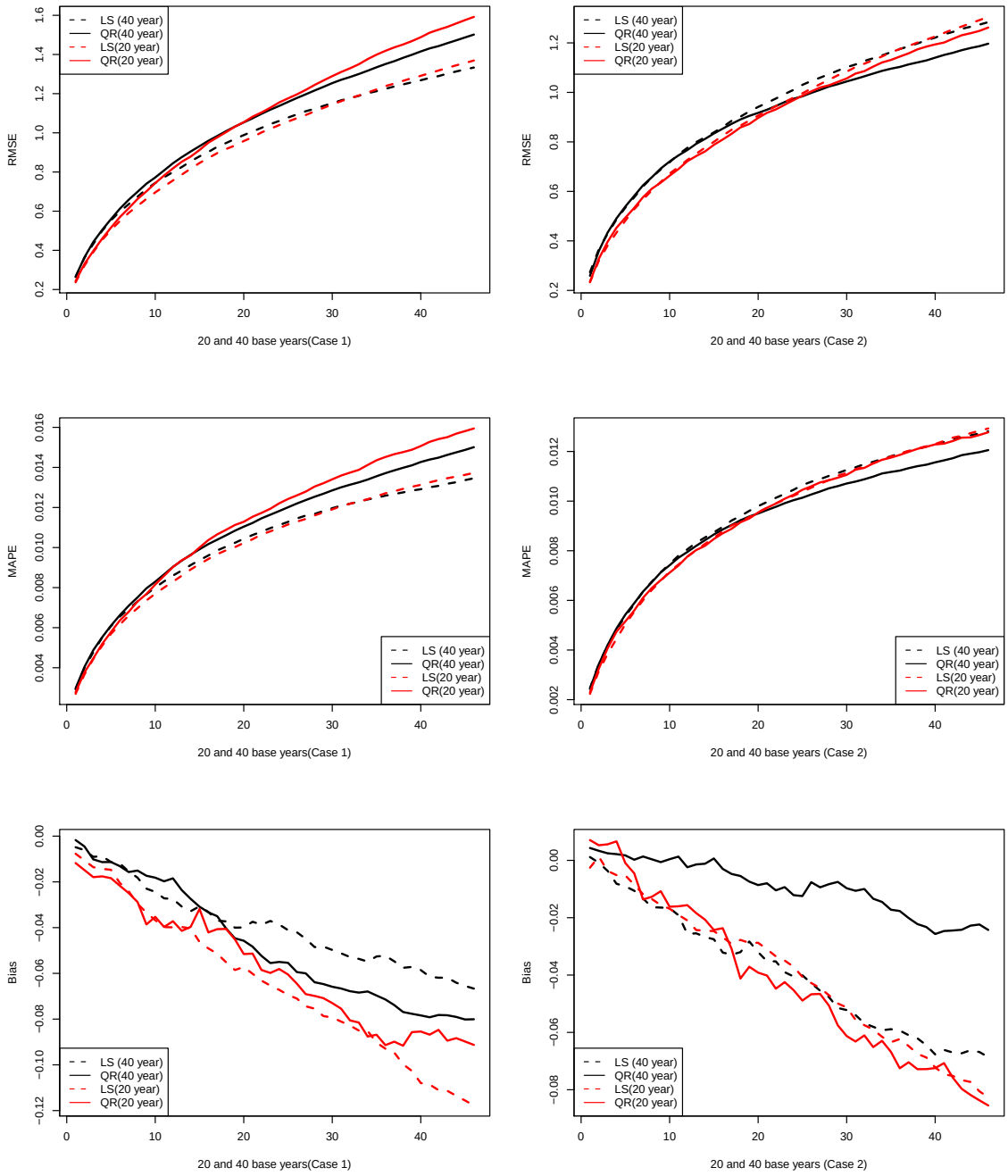

Figure 20: Simulation study on forecast performance based on LS and QR. The seed is obtained from *Canadian male* mortality data from 1921 to 2006. Forecast errors of RMSE, MAPE, and bias for $e_0(t)$ for 45-year horizons are presented for both Case 1 (left column panel) and Case 2 (right column panel).

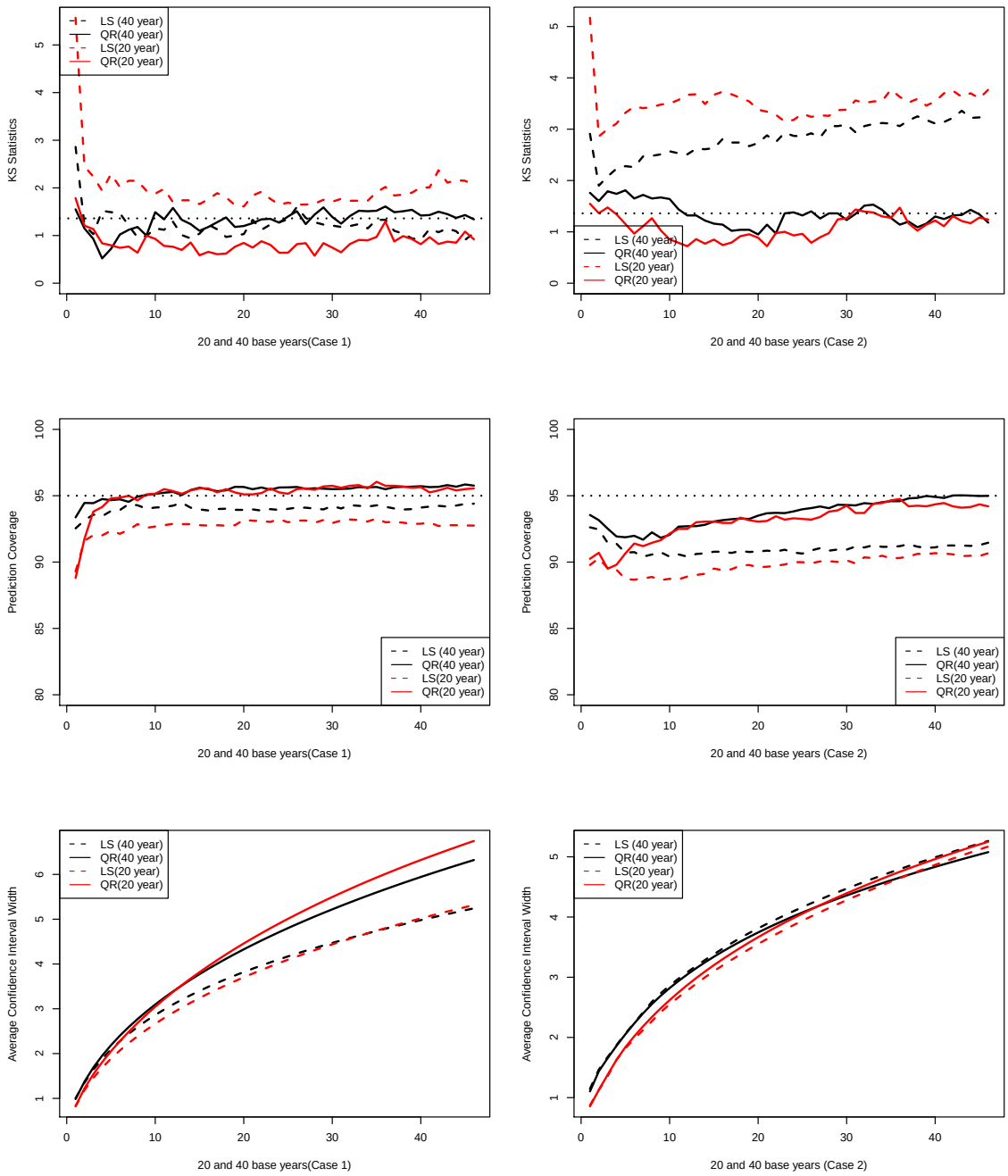

Figure 21: Simulation study on forecast performance based on LS and QR. The seed is obtained from *Canadian male* mortality data from 1921 to 2006. Kolmogorov-Smirnov statistic, coverage and average confidence interval width for $e_0(t)$ for 45-year horizons are presented for both Case 1 (left column panel) and Case 2 (right column panel). The critical value 1.36 from Kolmogorov-Smirnov test and the nominal coverage 95 percent are also marked.

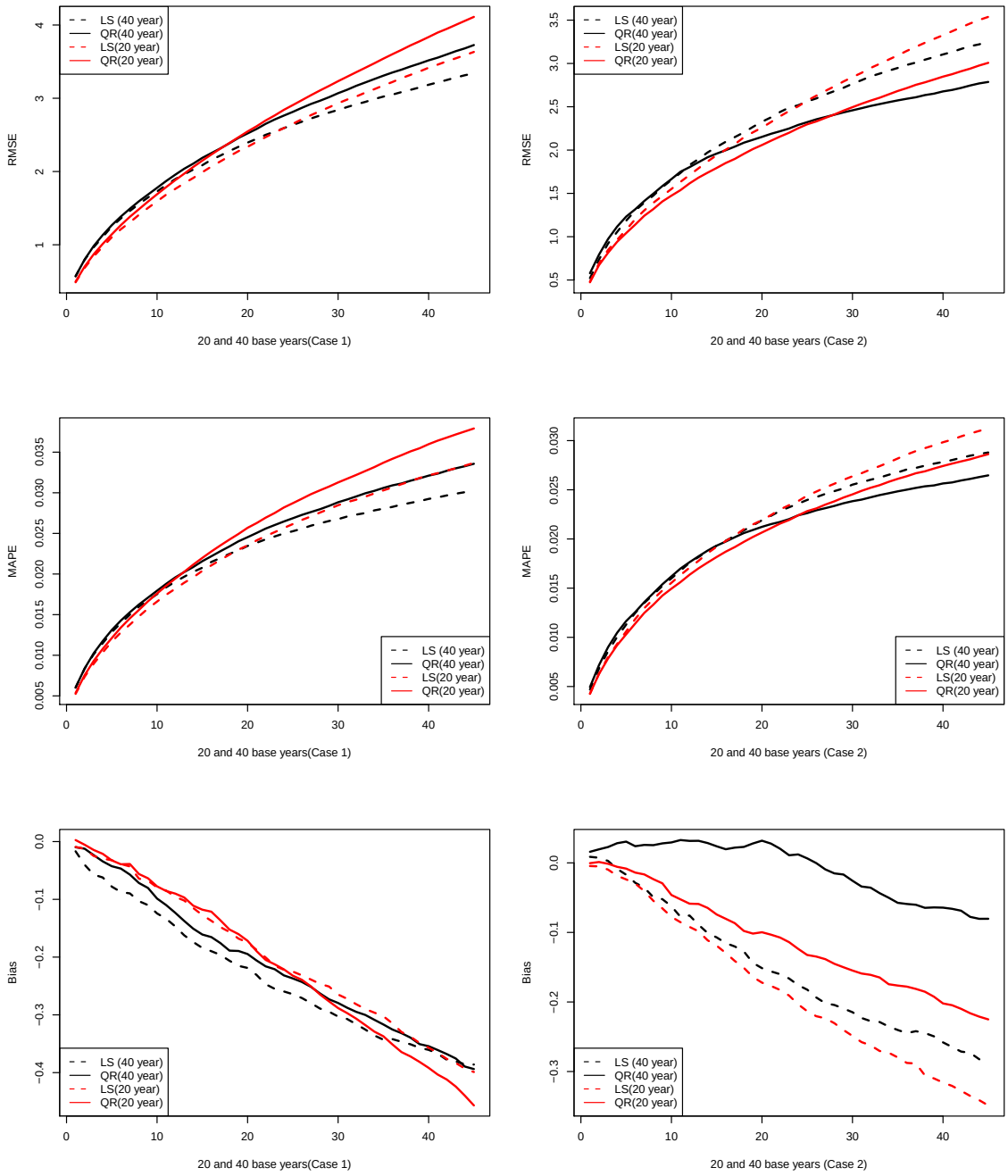

Figure 22: Simulation study on forecast performance based on LS and QR. The seed is obtained from *U.K. female* mortality data from 1922 to 2006. Forecast errors of RMSE, MAPE, and bias for $e_0(t)$ for 45-year horizons are presented for both Case 1 (left column panel) and Case 2 (right column panel).

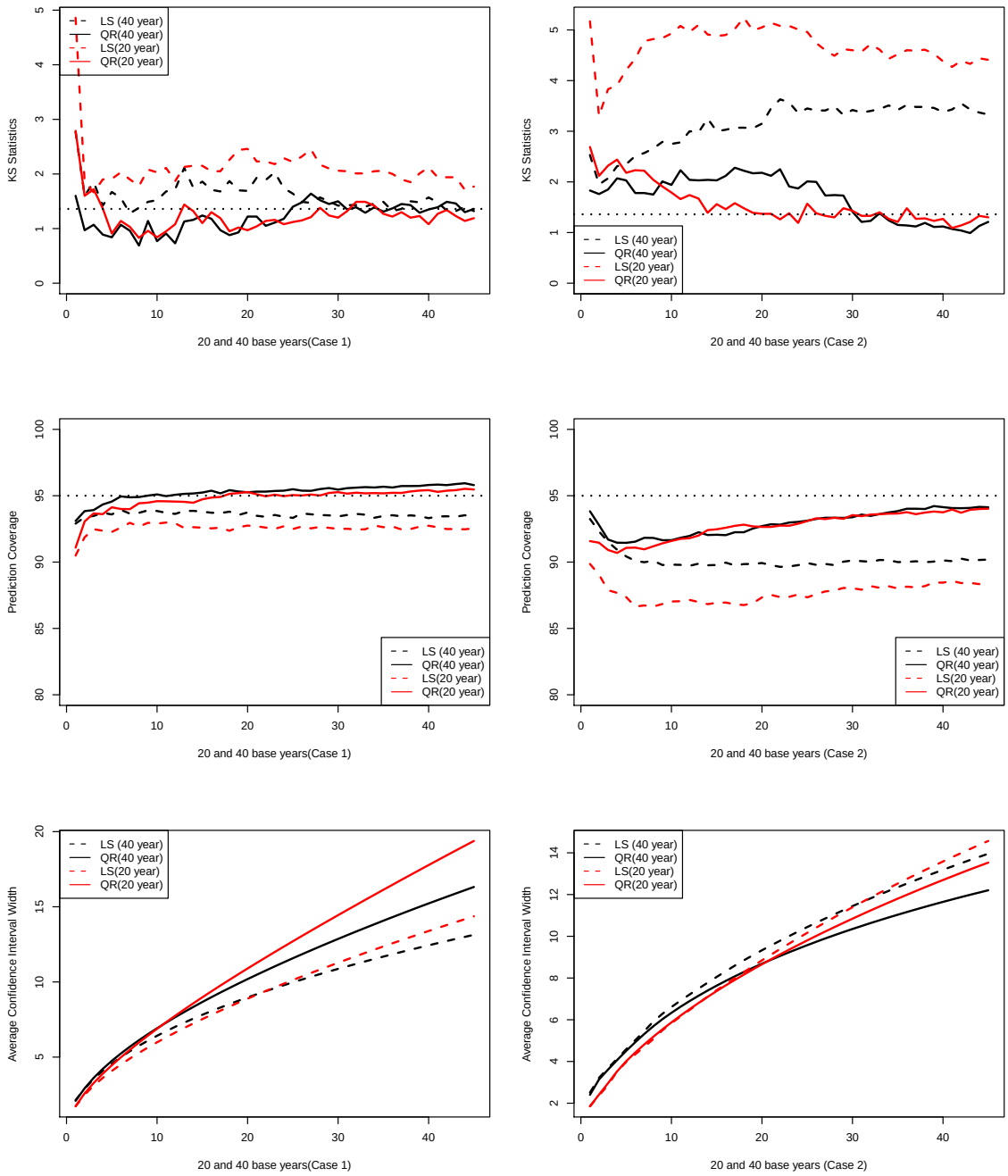

Figure 23: Simulation study on forecast performance based on LS and QR. The seed is obtained from *U.K. female* mortality data from 1922 to 2006. Kolmogorov-Smirnov statistic, coverage and average confidence interval width for $e_0(t)$ for 45-year horizons are presented for both Case 1 (left column panel) and Case 2 (right column panel). The critical value 1.36 from Kolmogorov-Smirnov test and the nominal coverage 95 percent are also marked.

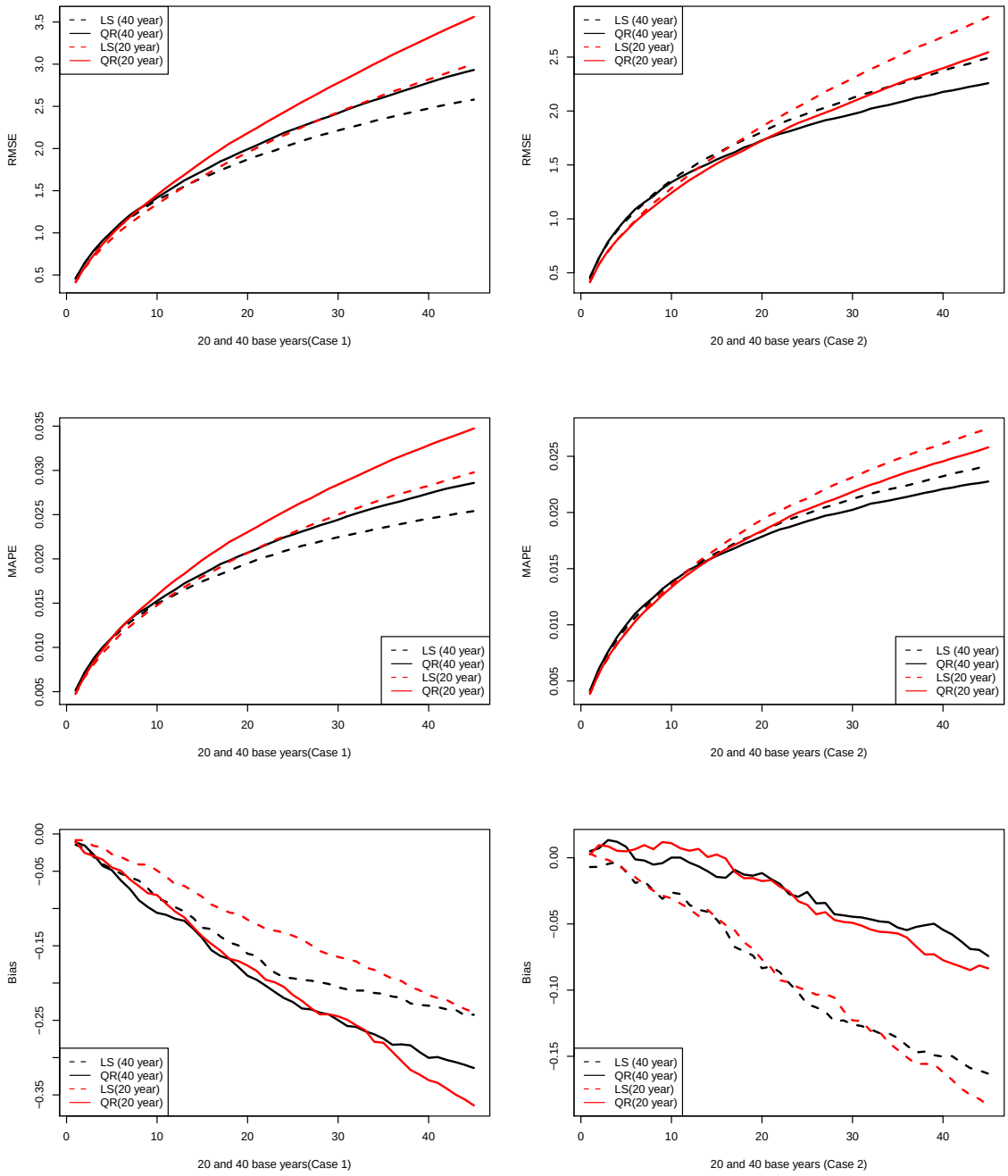

Figure 24: Simulation study on forecast performance based on LS and QR. The seed is obtained from *U.K. male* mortality data from 1922 to 2006. Forecast errors of RMSE, MAPE, and bias for $e_0(t)$ for 45-year horizons are presented for both Case 1 (left column panel) and Case 2 (right column panel).

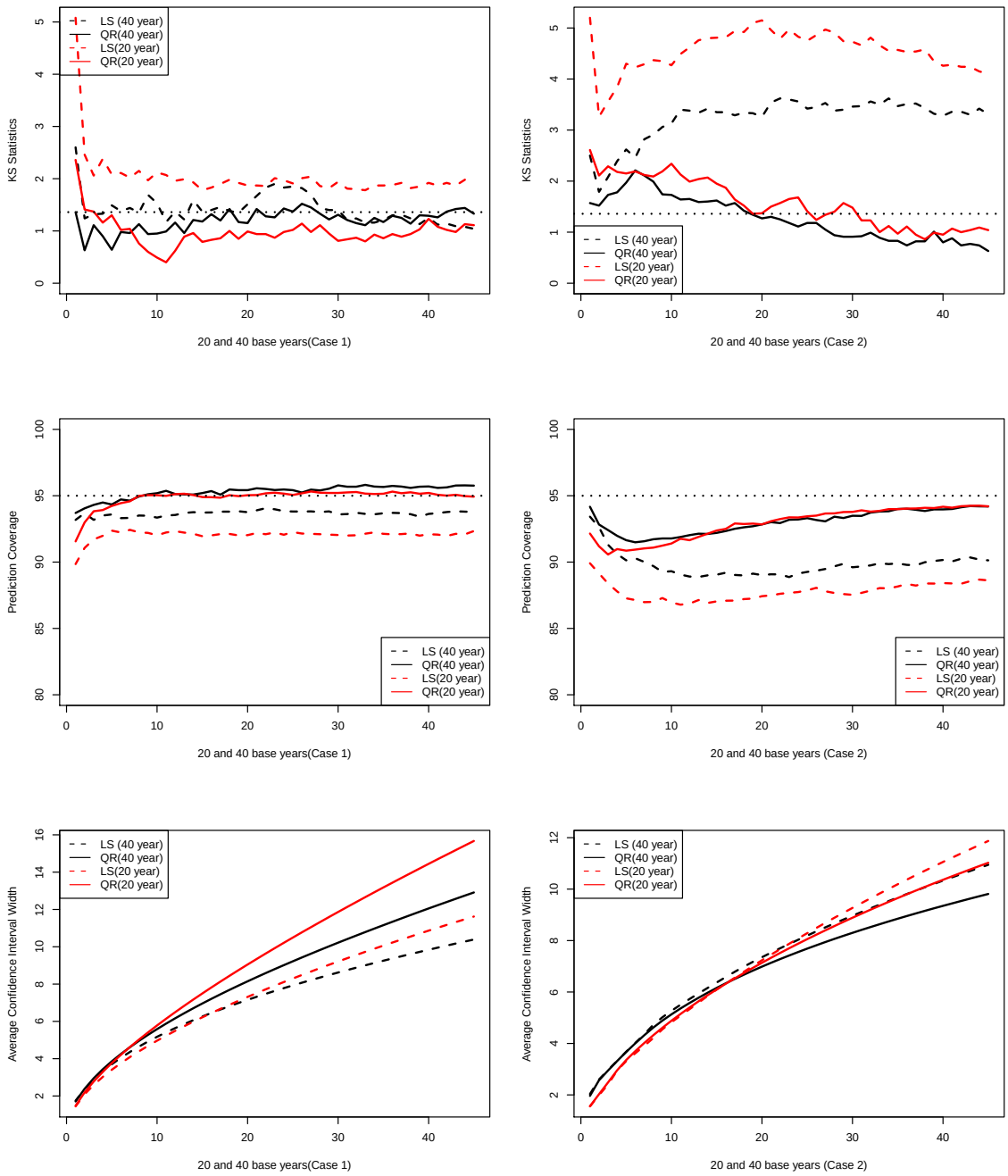

Figure 25: Simulation study on forecast performance based on LS and QR. The seed is obtained from *U.K. male* mortality data from 1922 to 2006. Kolmogorov-Smirnov statistic, coverage and average confidence interval width for $e_0(t)$ for 45-year horizons are presented for both Case 1 (left column panel) and Case 2 (right column panel). The critical value 1.36 from Kolmogorov-Smirnov test and the nominal coverage 95 percent are also marked.

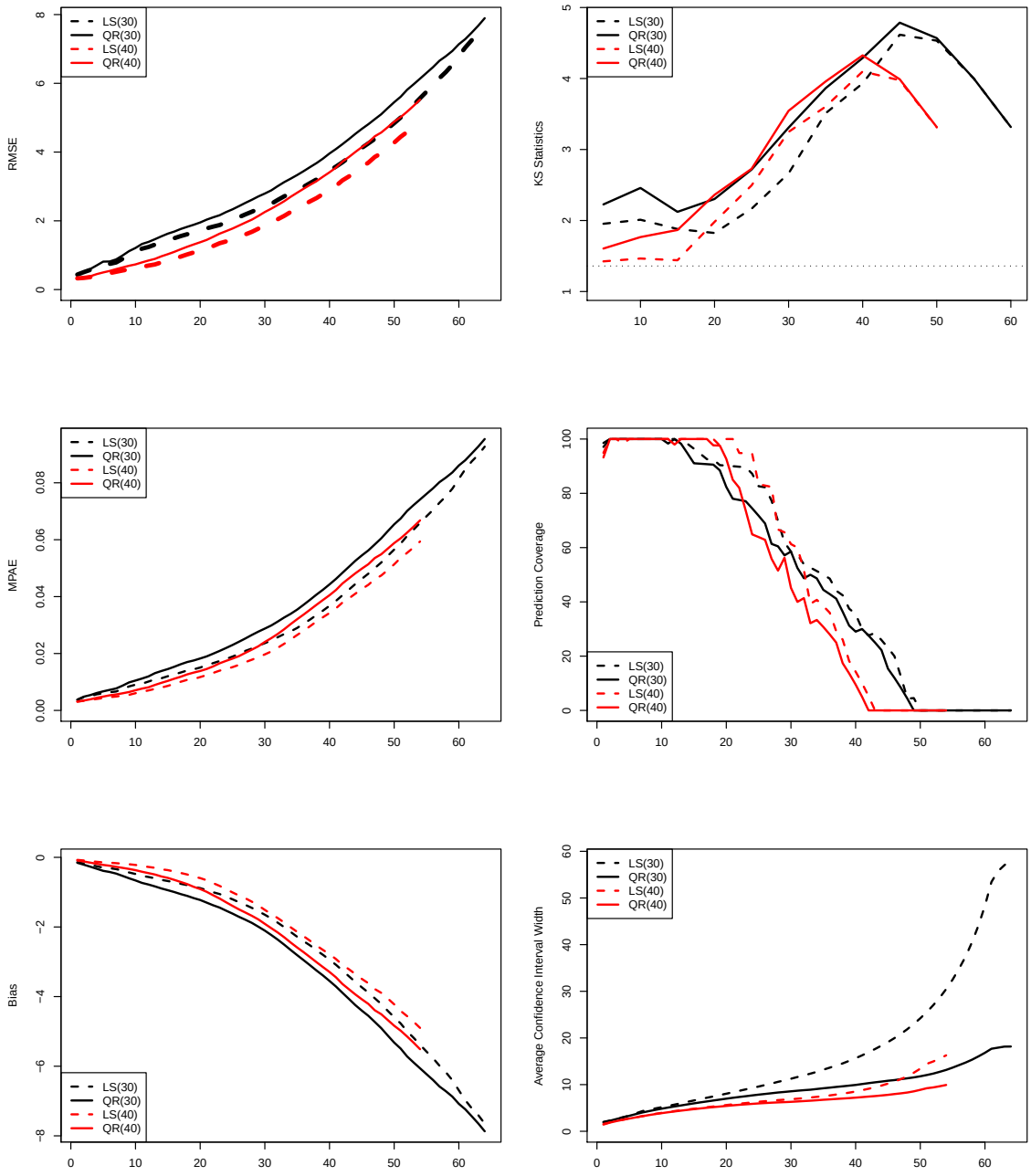

Figure 26: Real data analysis on *Swedish female* mortality data from 1921 to 2006 based on LS and QR using 30- or 40-year base period. Rolling window procedure is applied and the first jump off year is 1960. Presented are forecast criteria: RMSE, MAPE, bias, Kolmogorov-Smirnov statistic, coverage and average confidence interval width for $e_0(t)$.

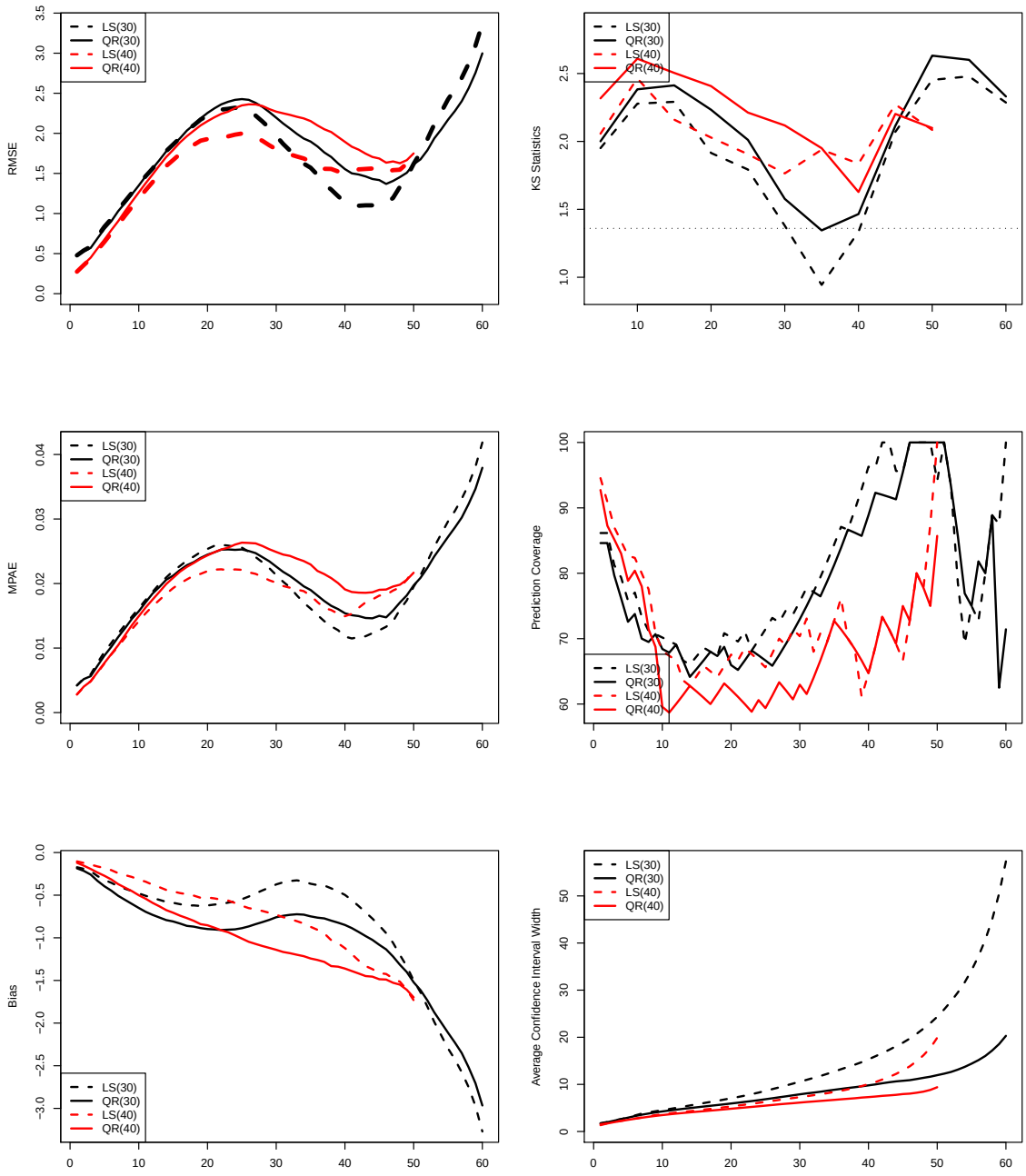

Figure 27: Real data analysis on *Swedish male* mortality data from 1907 to 2006 based on LS and QR using 30- or 40-year base period. Rolling window procedure is applied and the first jump off year is 1960. Presented are forecast criteria: RMSE, MAPE, bias, Kolmogorov-Smirnov statistic, coverage and average confidence interval width for $e_0(t)$.

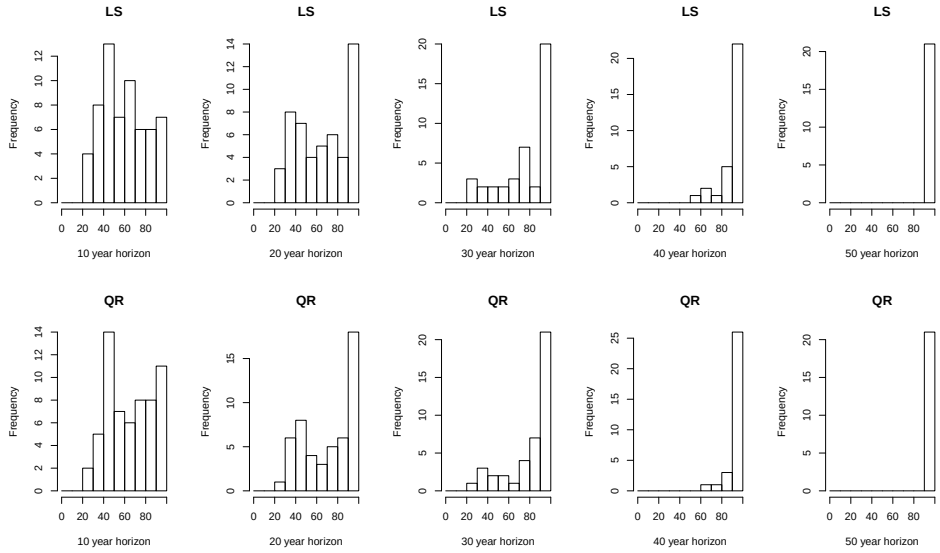

Figure 28: Real data analysis on *Swedish female* mortality data from 1907 to 2006. Rolling window procedure is applied, the base period is 30-year and the first jump off year is 1960. Presented is the percentile histograms from LS and QR

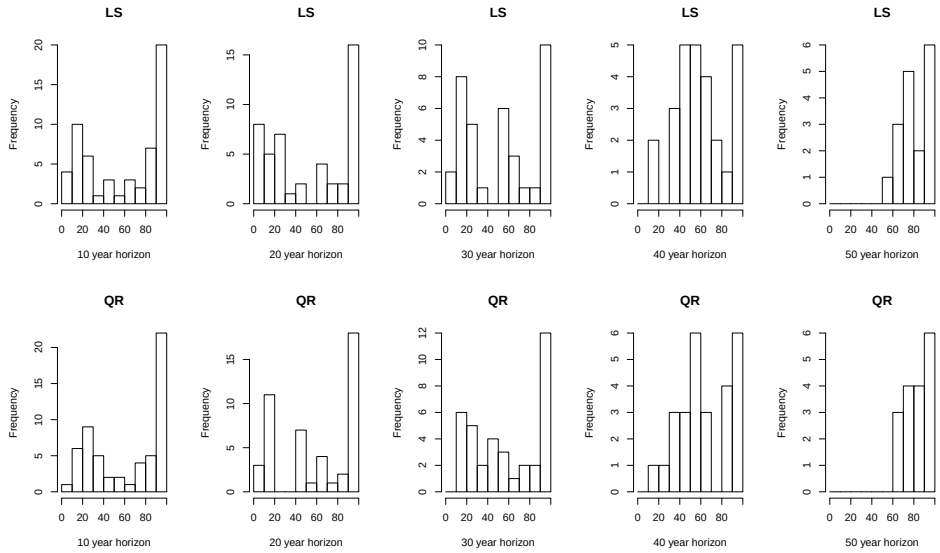

Figure 29: Real data analysis on *Swedish male* mortality data from 1907 to 2006. Rolling window procedure is applied, the base period is 30-year and the first jump off year is 1960. Presented is the percentile histograms from LS and QR

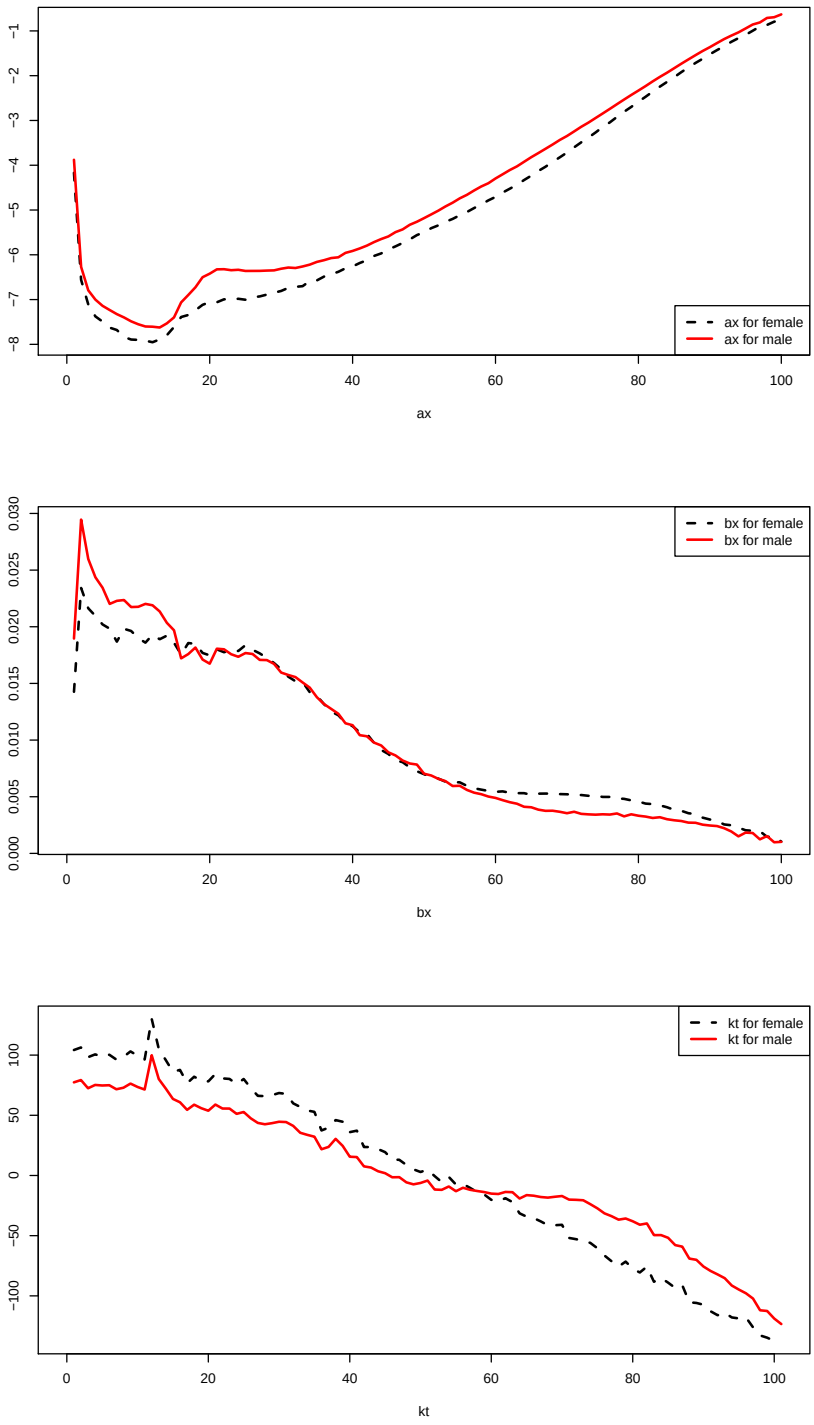

Figure 30: The fitted LC model estimates based on Swedish female and male mortality data from 1907 to 2006.

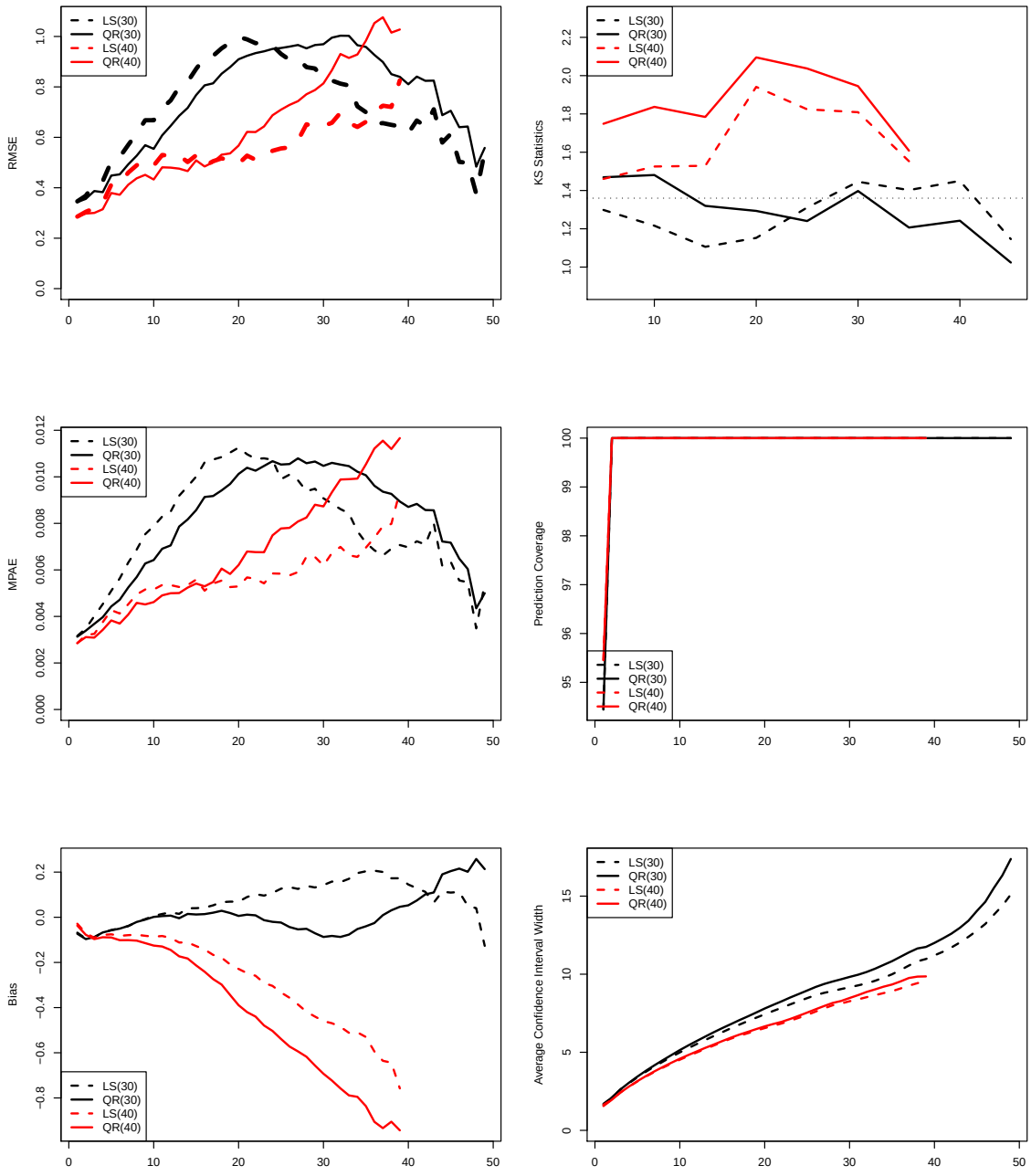

Figure 31: Real data analysis on *U.K. female* mortality data from 1921 to 2006 based on LS and QR using 30- or 40-year base period. Rolling window procedure is applied and the first jump off year is 1960. Presented are forecast criteria: RMSE, MAPE, bias, Kolmogorov-Smirnov statistic, coverage and average confidence interval width for $e_0(t)$.

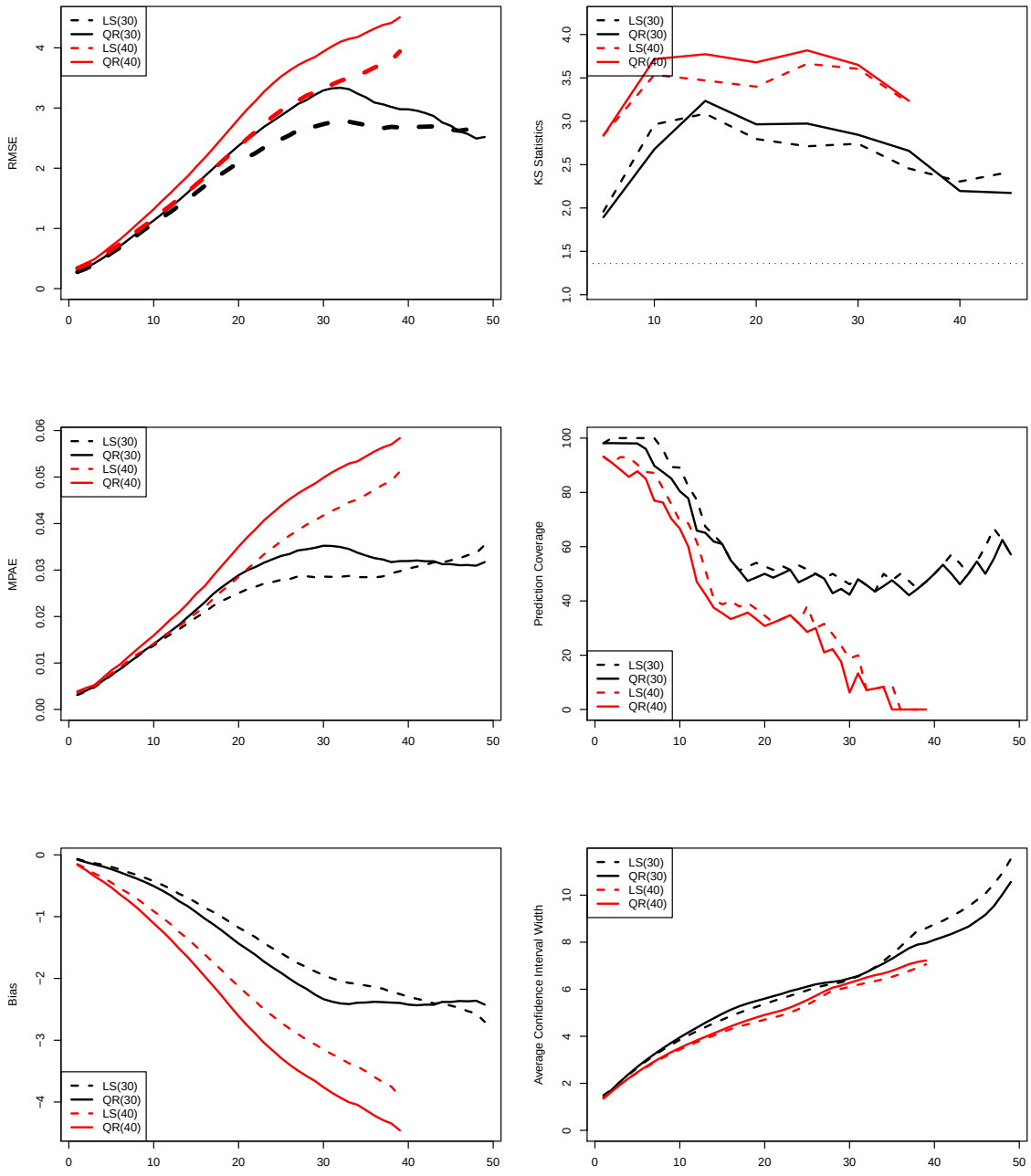

Figure 32: Real data analysis on *U.K. male* mortality data from 1921 to 2006 based on LS and QR using 30- or 40-year base period. Rolling window procedure is applied and the first jump off year is 1960. Presented are forecast criteria: RMSE, MAPE, bias, Kolmogorov-Smirnov statistic, coverage and average confidence interval width for $e_0(t)$.

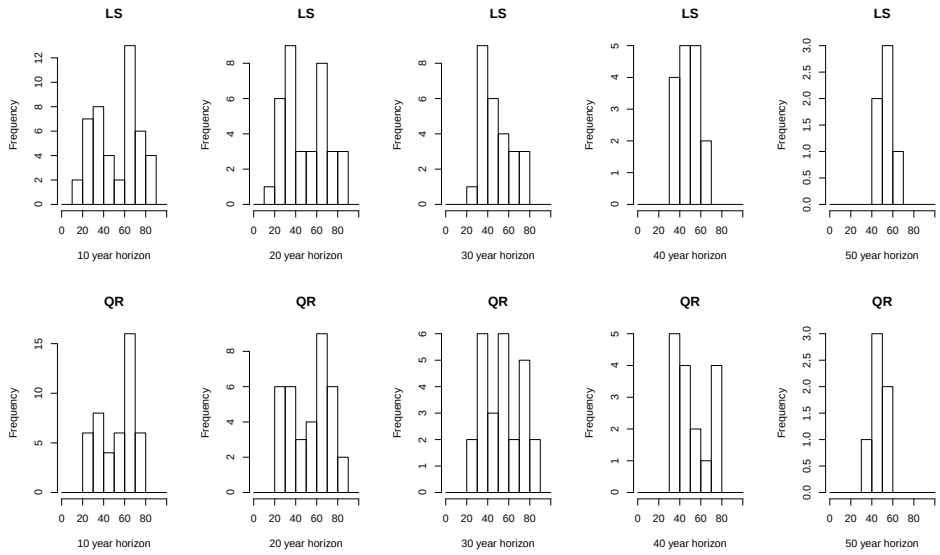

Figure 33: Real data analysis on *U.K. female* mortality data from 1921 to 2006. Rolling window procedure is applied, the base period is 30-year and the first jump off year is 1960. Presented is the percentile histograms from LS and QR

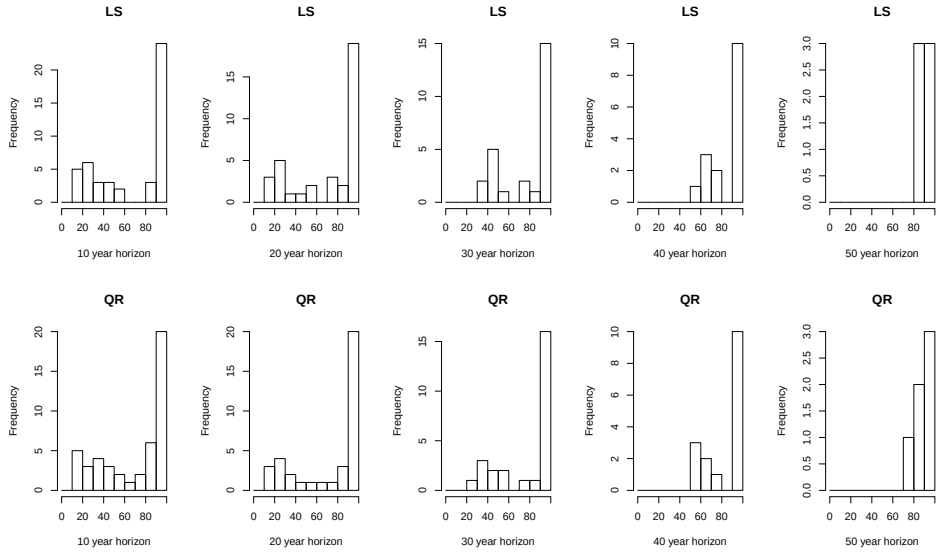

Figure 34: Real data analysis on *U.K. male* mortality data from 1921 to 2006. Rolling window procedure is applied, the base period is 30-year and the first jump off year is 1960. Presented is the percentile histograms from LS and QR

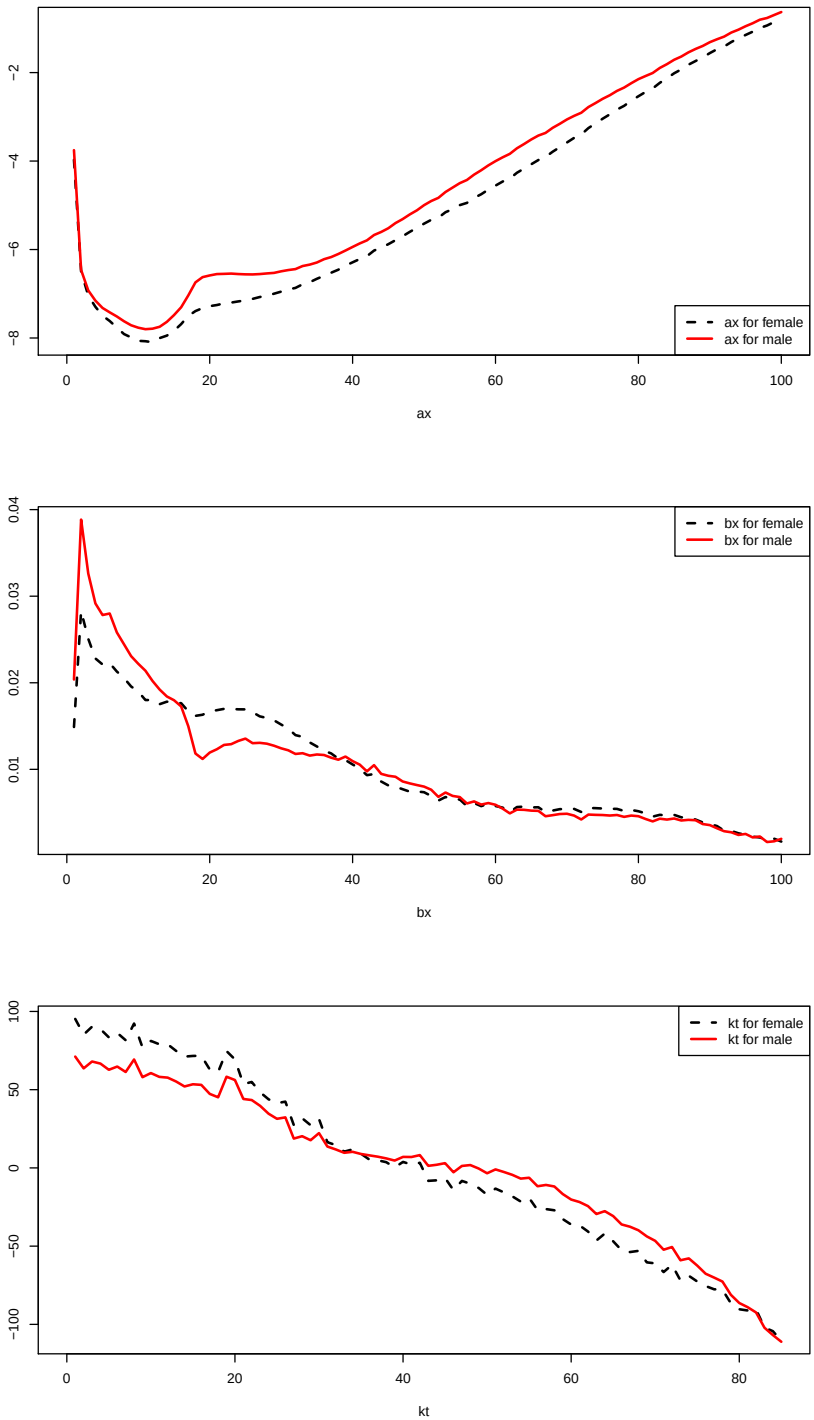

Figure 35: The fitted LC model estimates based on U.K. female and male mortality data from 1922 to 2006.