Approximate leave-future-out cross-validation for time series models

Paul-Christian Bürkner, Jonah Gabry, Aki Vehtari

Abstract

One of the most common goals of a time series analysis is to use the observed series to inform predictions for future observations. In the absence of any actual new data to predict, cross-validation can be used to measure a model's predictive accuracy for instance for the purpose of model comparison or selection. As exact cross-validation is often practically infeasible for Bayesian models because it requires too much time, approximate cross-validation methods have been developed; most notably methods for leave-one-out cross-validation (LOO-CV). However, for time-series models, it does not make sense to leave out observations one at a time because then we are allowing information from the future to influence predictions of the past. To apply the idea of cross-validation to time-series models, we thus need some form of leave-future-out cross-validation (LFO-CV). Like exact LOO-CV, exact LFO-CV requires refitting the model many times to different subsets of the data, which is computationally very costly for most nontrivial examples, in particular for Bayesian models. Using Pareto-smoothed importance sampling, we propose a method for approximating exact LFO-CV that drastically reduces the computational burden while also providing informative diagnostics about the quality of the approximation.

1 Introduction

A time series is a set of observations each one being recorded at a specific time (Brockwell et al., 2002). In statistics, a wide range of time series models has been developed, which find application in nearly all empirical sciences (e.g., see Brockwell et al., 2002; Hamilton, 1994). One of the most common goals of a time series analysis is to use the observed series to inform predictions for future observations. When working with discrete time series – in which time points form discrete set – we will refer to the task of predicting a sequence of M future observations as M-step-ahead prediction (M-SAP). Fortunately, once we have fit a Bayesian model and can sample from the posterior predictive distribution, it is straightforward to generate predictions as far into the future as we want. It is also straightforward to evaluate the M-SAP performance of a time series model by comparing the predictions to the observed sequence of M future data points once they become available.

Unfortunately, we are often in the position of having to use a model to inform decisions *before* we can collect the future observations required for assessing the predictive performance. If we have many competing models we may also need to first decide which of the models (or which combination of the models) we should rely on for predictions (Geisser and Eddy, 1979; Hoeting et al., 1999; Vehtari and Lampinen, 2002; Ando and Tsay, 2010; Vehtari and Ojanen, 2012). In the absence of new data with which to evaluate predictive performance, one general approach for evaluating a model's predictive accuracy is cross-validation (Vehtari and Lampinen, 2002). When doing cross-validation, the data is split into two subsets. Based on the first subset we fit the statistical model and then evaluate its predictive accuracy for the second subset. We may do this once or many times, each time leaving out another subset.

If there were no time dependence in the data or if the focus is to assess the non-time-dependent part of the model, we could use methods like leave-one-out cross-validation (LOO-CV). For a data set with N observations, we refit the model N times, each time leaving out one of the N observations and assessing how well the model predicts the left-out observation. LOO-CV is very expensive computationally in most realistic settings, but the Pareto smoothed importance sampling (PSIS; Vehtari et al., 2017b,a) algorithm allows for approximating exact LOO-CV with PSIS-LOO-CV. PSIS-LOO-CV requires only a single fit of the full model and comes with diagnostics for assessing the validity of the approximation.

With a time series we can do something similar to LOO-CV but, except in a few cases, it does not make sense to leave out observations one at a time because then we are allowing information from the future to influence predictions of the past (i.e., times $t+1, t+2, \ldots$ should not be used to predict for time t). To apply the idea of cross-validation to the M-SAP case, instead of leave-*one*-out cross-validation we need some form of leave-*future*-out cross-validation (LFO-CV). As we will demonstrate in this paper, LFO-CV does not refer to one particular prediction task but rather to various possible cross-validation approaches that all involve some form of prediction for new time series data. Like exact LOO-CV, exact LFO-CV requires refitting the model many times to different subsets of the data, which is computationally very costly for most nontrivial examples, in particular for Bayesian analyses where refitting the model means estimating a new posterior distribution rather than a point estimate.

Although PSIS-LOO-CV provides an efficient approximation to exact LOO-CV, until now there has not been an analogous approximation to exact LFO-CV that drastically reduces the computational burden while also providing informative diagnostics about the quality of the approximation. In this paper we present PSIS-LFO-CV, an algorithm that typically only requires refitting the time-series model a small number times and will make LFO-CV tractable for many more realistic applications than previously possible.

The structure of the paper is as follows. In Section 2 we introduce the idea and various forms of M-step-ahead predictions and how to approximate it using PSIS. In Section 3, we evaluate the accuarancy of the approximation using extensive simulations, before we provide two real world case studies about the change of level of Lake Huron and the day of the cherry blossoms in Japan across the years in Section 4. We end with a discussion of the usefulness and limitations of the approach in Section 5.

2 M-step-ahead predictions

Assume we have a time series of observations $y = (y_1, y_2, \ldots, y_N)$ and let L be the *minimum* number of observations from the series that we will require before making predictions for future data. Depending on the application and how informative the data is, it may not be possible to make reasonable predictions for y_i based on $(y_1, \ldots, y_{i-1})$ until i is large enough so that we can learn enough about the time series to predict future observations. Setting $L = 10$, for example, means that we will only assess predictive performance starting with observation y_{11}, so that we always have at least 10 previous observations to condition on.

In order to assess M-SAP performance we would like to compute the predictive densities

$$p(y_{i<M} \,|\, y_{<i}) = p(y_i, \ldots, y_{i+M-1} \,|\, y_1, ..., y_{i-1}) \tag{1}$$

for each $i \in \{L+1, \ldots, N-M+1\}$, where we use $y_{i<M} = (y_i, \ldots, y_{i+M-1})$ and $y_{<i} = (y_1, \ldots, y_{i-1})$ to shorten the notation. As a global measure of predictive accuracy, we can use the expected log posterior density (ELPD), which, for LFO-CV, can be defined as follows:

$$\mathrm{ELPD} = \sum_{i=L+1}^{N-M+1} \log p(y_{i<M} \,|\, y_{<i}) \tag{2}$$

The quantities $p(y_{i<M} \,|\, y_{<i})$ can be computed with the help of the posterior distribution $p(\theta \,|\, y_{<i})$ of the parameters θ conditional on only the first $i-1$ observations of the time-series:

$$p(y_{i<M} \,|\, y_{<i}) = \int p(y_{i<M} \,|\, y_{<i}, \theta)\, p(\theta \,|\, y_{<i})\, d\theta. \tag{3}$$

Having obtained S draws $(\theta_{<i}^{(1)}, \ldots, \theta_{<i}^{(S)})$ from the posterior distribution $p(\theta \,|\, y_{<i})$, we can estimate $p(y_{i<M}|y_{<i})$ as

$$p(y_{i<M} \,|\, y_{<i}) \approx \frac{1}{S} \sum_{s=1}^{S} p(y_{i<M} \,|\, y_{<i}, \theta_{<i}^{(s)}). \tag{4}$$

If we deal with factorizable models in which the response values are conditionally independent given the parameters, the likelihood can be written in the familiar form

$$p(y \,|\, \theta) = \prod_{n=1}^{N} p(y_j \,|\, \theta). \tag{5}$$

In this case, $p(y_{i<M} \,|\, y_{<i}, \theta_{<i})$ reduces to

$$p(y_{i<M} \,|\, \theta_{<i}) = \prod_{n=i}^{i+M-1} p(y_j \,|\, \theta_{<i}), \tag{6}$$

due to the assumption of conditional independence between $y_{i<M}$ and $y_{<i}$ given $\theta_{<i}$. Non-factorizable models, which do not make this assumption, are discussed in more detail in Bürkner et al. (view).

2.1 Approximate M-step-ahead predictions

Unfortunately, the math above makes use of the posterior distributions from many different fits of the model to different subsets of the data. That is, to obtain the predictive density $p(y_{i<M} \,|\, y_{<i})$ requires fitting a model to only the first $i-1$ data points, and we will need to do this for every value of i under consideration (all $i \in \{L+1, \ldots, N-M+1\}$).

To reduce the number of models that need to be fit for the purpose of obtaining each of the densities $p(y_{i<M} \,|\, y_{<i})$, we propose the following algorithm. Starting with $i = N-M+1$, we approximate each $p(y_{i<M} \,|\, y_{<i})$ using Pareto smoothed importance sampling (PSIS; Vehtari et al., 2017b,a):

$$
p(y_{i<M} \mid y_{<i}) \approx \frac{\sum_{s=1}^{S} w_i^{(s)} \, p(y_{i<M} \mid \theta^{(s)})}{\sum_{s=1}^{S} w_i^{(s)}}, \tag{7}
$$

where $w_i^{(s)}$ are importance weights and $\theta^{(s)}$ are draws from the posterior distribution based on *all* observations. To obtain $w_i^{(s)}$, we first compute the raw importance ratios

$$
r_i^{(s)} \propto \frac{1}{\prod_{j \in J_i} p(y_j \mid \theta^{(s)})}, \tag{8}
$$

and then stabilize them using PSIS as described in Vehtari et al. (2017a). The index set J_i contains all the indices of observations which are part of the actually fitted model but not of the model whose predictive performance we are trying to approximate. That is, for the starting value $i = N - M + 1$, we have $J_i = \{i, \ldots, N\}$. This approach to computing importance ratios is a generalization of the approach used in PSIS-LOO-CV, where only a single observation is left out at a time and thus $J_i = i$ for all i.

Starting from $i = N - M + 1$, we gradually *decrease* i by 1 (i.e., we move backwards in time) and repeat the process. At some observation i, the variability of importance ratios $r_i^{(s)}$ will become too large and importance sampling fails. We will refer to this particular value of i as $i_1^\star$. To identify the value of $i_1^\star$, we check for which value of i does the estimated shape parameter k of the generalized Pareto distribution first cross a certain threshold τ (Vehtari et al., 2017a). Only then do we refit the model using only observations before $i_1^\star$ and then restart the process. Until the next refit, we have $J_i = \{i, \ldots, i_1^\star - 1\}$ for $i < i_1^\star$, as the refitted model only contains the observations up to index $N_1^\star = i_1^\star - 1$. An illustration of the above described procedure is shown in Figure 1.

In some cases we may only need to refit once and in other cases we will find a value $i_2^\star$ that requires a second refitting, maybe an $i_3^\star$ that requires a third refitting, and so on. We repeat the refitting as few times as is required (only if $k > \tau$) until we arrive at $i = L + 1$. Recall that L is the minimum number of observations we have deemed acceptable for making predictions (setting $L = 0$ means predictions of all observations should be computed).

The threshold τ is crucial to the accuracy and speed of the proposed algorithm. If τ is too large, we need fewer refits and thus achieve higher speed, but accuracy is likely to suffer. If τ is too small, we get high accuracy but a lot of refits to that speed will drop noticably. When performing exact CV of Bayesian models, almost all of the computational time is spend fitting models, while the time needed to do predictions is negligeble in comparison. That is, a reduction of the number of refits basically implies a proportional reduction in the overall time necessary for CV of Bayesian models.

A mathematical analysis of the Pareto distribution reveals that approximate CV via PSIS is very likely to be highly accuracte as long as $k < 0.5$ (Vehtari et al., 2017a). In practice, PSIS-LOO-CV turned out to be robust for $k < 0.7$ (Vehtari et al., 2017b). That is, for PSIS-LFO-CV introduced in the present paper, we can expect an appropriate threshold to be somewhere between $0.5 \leq \tau \leq 0.7$. It is unlikely to be as high as $\tau = 0.7$, as the error made in the prediction of a certain observation i will propagate to the predictions of observations $i-1, i-2, \ldots$ until a refit is performed. That is, problematic observations with high k are likely to have stronger effects in LFO-CV than LOO-CV. We will come back to the issue of setting approriate thresholds in Section 3.

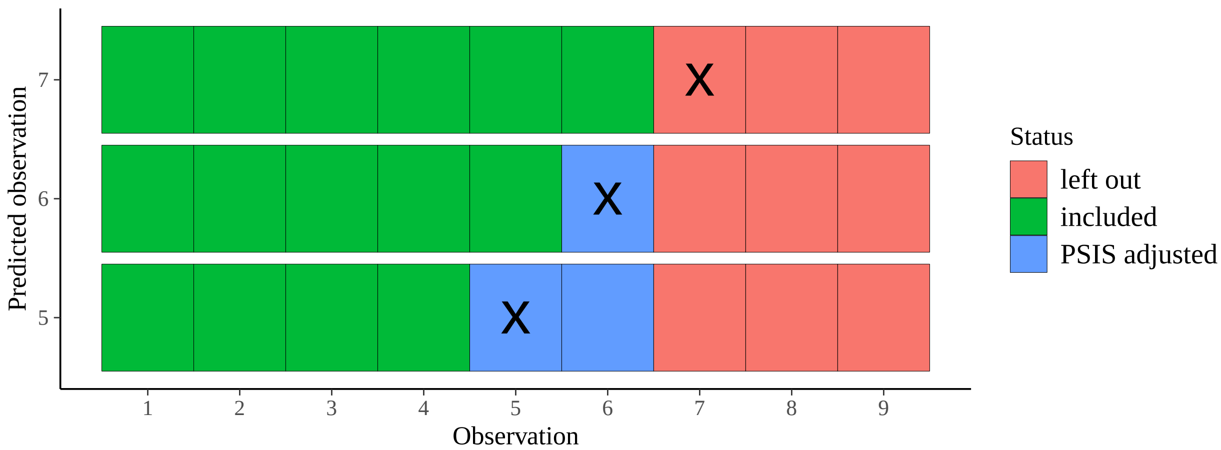

Figure 1: Visualisation of PSIS approximated one-step-ahead predicitions. The `x` represents the currently predicted observation. In the shown example, the model was last refit at the 7th observation.

2.2 Block M-step-ahead predictions

Depending on the particular time-series data and model, the Pareto k estimates may exceed τ rather quickly (i.e., after only few observations) and so a lot of refits may be required even when carrying out the PSIS approximation to LFO-CV. In this case, another option is to exclude only the block of B future values that directly follow the observations to be predicted while retaining all of the more distant values $y_{i>B} = (y_{i+B}, \ldots, y_N)$. This will usually result in lower Pareto k estimates and thus less refitting, but crucially alters the underlying prediction task, to which we will refer to as block-M-SAP.

The block-M-SAP version closely resembles the basic M-SAP only if values in the distant future, $y_{>B}$, contain little information about the current observations being predicted, apart from just increasing precision of the estimated parameters. Whether this assumption is justified will depend on the data and model. That is, if the time-series is non-stationary, distant future value will inform overall trends in the data and thus clearly inform predictions of the current observations being left-out. As a result, block-LFO-CV is only recommended for stationary time-series and corrsponding models.

There are more complexities that arise in block-M-SAP that we did not have to care about in standard M-SAP. One is that, by just removing the block, the time-series effectively gets split into two parts, one before and one after the block. This poses no problem for conditionally independent time-series models, where predictions just depend on the parameters and not on the former values of the time-series itself. However, if the model's predictions are *not* conditionally independent as is the case, for instance, in autoregressive models (see Section 3), the observations of the left-out block have to be modeled as missing values in order to retain the integreity of the time-series' predictions after the block. A related example from spatial statistics, in which the modeling of missing values is required for valid inference, can be found in Bürkner et al. (view).

Another complexity concerns the PSIS approximation of block-LFO-CV: Not only does the approximating model contain more observations than the current model whose predictions we are approximating, but it also may *not* contain observations that are present in the actual model. The latter observations are those right after the currently left-out block, which are included in the current model, but not in the approximating model as they were part of the block at the time the approximating model was (re-)fit. A visualitation of this situation is provided in Figure X. More formally, let $\overline{J}_i$ be the index set of observations that are missing in the approximating model at the time of predicting observation i. We find

$$\overline{J}_i = \{\max(i + B, N^\star + 1), \ldots, \min(N^\star + B, N)\} \tag{9}$$

if $\max(i+B, N^\star+1) \leq \min(N^\star+B, N)$ and $\overline{J}_i = \emptyset$ otherwise. As above, $N^\star$ refers to the largest observation included in the model fitting, that is $N^\star = i^\star - 1$ where i^* is the index of the latest refit. The raw importance ratios $r_i^{(s)}$ for each posterior draw s are then computed as

$$r_i^{(s)} \propto \frac{\prod_{j \in \overline{J}_i} p(y_j \mid \theta^{(s)})}{\prod_{j \in J_i} p(y_j \mid \theta^{(s)})} \tag{10}$$

before they are stabilized and further processed using PSIS (see Section 2.1).

3 Simulations

To evaluate the goodness of the approximation of PSIS-LFO-CV, we performed a simulation study by systematically varying the following conditions: The number M of future observations to be predicted took on values of $M = 1$ and $M = 4$. The number of future values to be excluded in the model fitting took on values of $B = \infty$ (i.e., leaving out the whole future), or $B = 10$ (i.e. leaving out only a block of 10 observations). The threshold τ of the Pareto k estimates was varied between $k = 0.5$ to $k = 0.7$ in steps of 0.1. In addition, we evaluated six different data generating models with linear and/or quadatric terms and/or autoregressive terms of order 2 (see Table X for an overview). These models are also illustrated graphically in Figure 2. In all conditions, the time-series consistent of $N = 200$ observations and the minimal number of observations to make predictions was set to $L = 25$ throughout.

Autoregressive (AR) models are some of the most commonly used time-series models. An AR(p) model – an autoregressive model of order p – can be defined as

$$y_i = \eta_i + \sum_{k=1}^{p} \varphi_k y_{i-k} + \varepsilon_i, \tag{11}$$

where η_i is the linear predictor for the ith observation, ϕ_k are the autoregressive parameters and ε_i are pairwise independent errors, which are usually assumed to be normally distributed with equal variance σ^2. The model implies a recursive formula that allows for computing the right-hand side of the above equation for observation i based on the values of the equations for previous observations. Thus, by definition, responses of AR-models are not conditionally independent. However they are still factorizable, that is we may write down a separate likelihood contribution per observation (see Bürkner et al., view, for more discussion on factorizability of statistical models).

All simulations were done in R (R Core Team, 2018) using the brms package (Bürkner, 2017, 2018) together with the probabilistic programming language Stan (Carpenter et al., 2017) for the modeling fitting, the loo

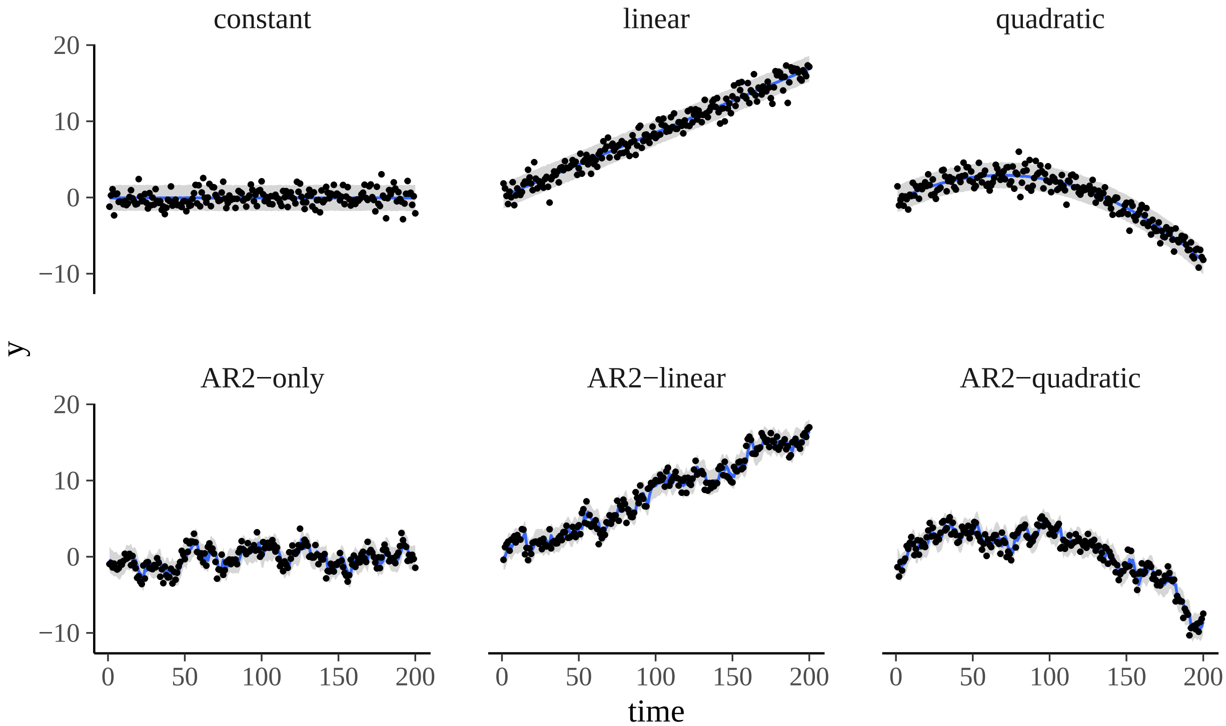

Figure 2: Illustration of the models used in the simulations.

package (Vehtari et al., 2017b) for the PSIS computation, and several tidyverse packages (Wickham, 2017) for data processing. The full code used in the simulations as well as all results are available on (add link).

3.1 Results

Results of the 1-SAP simulations leaving out all future values are visualized in Figure 3. Comparing the columns of Figure 3, it is clearly visible that the accuracy of the PSIS approximation increases with decreasing τ, up to almost perfect accuracy for $\tau = 0.5$. At the same time, the percentage of observations at which refitting the model was required increased substantially with decreasing τ (see Table 1). Using $\tau = 0.6$ induced a slight positive bias in PSIS-LFO-CV, but also reduced the number of required refits by roughly 30%. Another 30% reduction in the number of refits was achieved by using $\tau = 0.7$ but at the cost of disproportionally increasing the positive bias in PSIS-LFO-CV. Further, PSIS-LOO-CV turned out to be a bad measure of 1-SAP performance when leaving out all future values for all non-constant models in particular those with a trend in the time-series (see light-blue histograms in Figure 3).

Results of the 4-SAP simulations leaving out all future values are visualized in Figure 4. Comparing the columns of Figure 4, it is clearly visible that the accuracy of the PSIS approximation increases with decreasing τ, up to almost perfect accuracy for $\tau = 0.5$. At the same time, the percentage of observations at which refitting the model was required increased substantially with decreasing τ (see Table 1). In light of the corresponding 1-SAP results (see above), this is not surprising as the procedure to determining the necessity of a refit is independent of M (see Section 2.1). Using $\tau = 0.6$ again induced a slight positive bias in

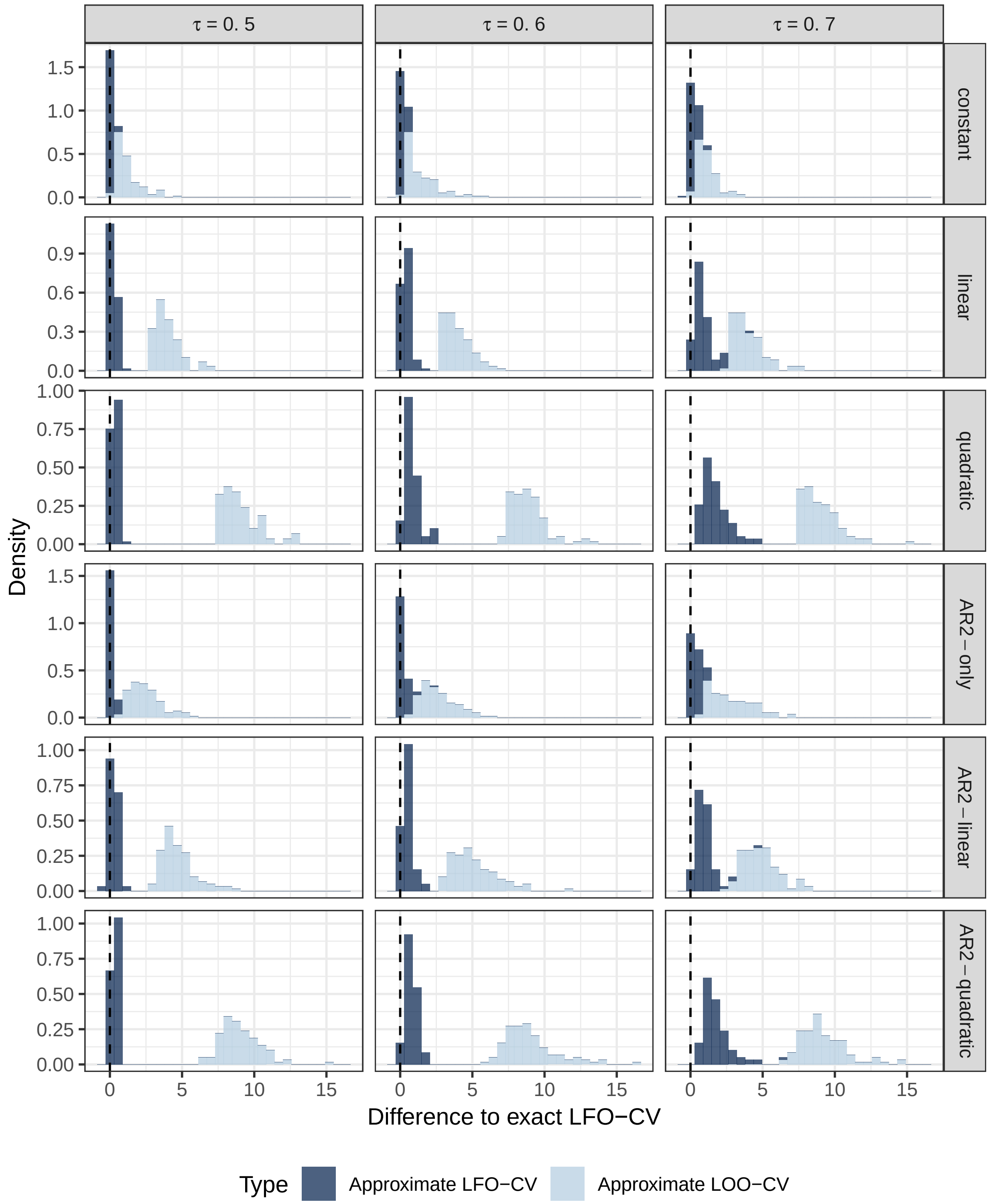

Figure 3: Simulation results of 1-step-ahead predictions.

Table 1: Mean percentages of required refits.

B	M	τ	constant	linear	quadratic	AR2-only	AR2-linear	AR2-quadratic
∞	1	0.5	0.03	0.08	0.17	0.05	0.09	0.18
		0.6	0.02	0.06	0.12	0.03	0.06	0.13
		0.7	0.01	0.04	0.08	0.02	0.04	0.08
	4	0.5	0.03	0.08	0.17	0.05	0.09	0.18
		0.6	0.02	0.06	0.12	0.03	0.06	0.13
		0.7	0.01	0.04	0.09	0.02	0.04	0.08
10	1	0.5	0.00	0.00	0.01	0.01	0.01	0.02
		0.6	0.00	0.00	0.00	0.00	0.00	0.01
		0.7	0.00	0.00	0.00	0.00	0.00	0.00
	4	0.5	0.00	0.00	0.01	0.01	0.01	0.02
		0.6	0.00	0.00	0.00	0.00	0.00	0.01
		0.7	0.00	0.00	0.00	0.00	0.00	0.00

Note: Results are based on 100 simulation trials of time-series with $N = 200$ observations requiring at least $L = 25$ observations to make predictions. Abbreviations: tau = threshold of the Pareto-k-estimates; M = number of predicted future observations; B = number of left-out future observations.

PSIS-LFO-CV, but also reduced the number of required refits by roughly 30%. Another 30% reduction in the number of refits was achieved by using $\tau = 0.7$ but at the cost of disproportionally increasing the positive bias in PSIS-LFO-CV. PSIS-LOO-CV is not displayed in Figure 4 as the number of observations predicted as each step (4 vs. 1) renders 4-SAP LFO-CV and LOO-CV incomparable.

Results of the block-1-SAP simulations leaving out a block of $B = 10$ future values are visualized in Figure 5. It can be seen that, in this case, PSIS-LFO-CV provides an closed to unbiased approximation of the corresponding exact LFO-CV for all investigated conditions, that is regardless of the threshold τ or the data generating model. The number of required refits was not only much smaller than when leaving out all future values, but practically approched zero for most conditions. Notably, PSIS-LOO-CV was similarily accurate than PSIS-LFO-CV as indicated by the strongly overlapping histograms in Figure 5. This is plausible given that LOO-CV and LFO-CV of block-1-SAP only differ in whether they include the relatively few observations in the block when fitting the approximating model.

Results of the block-4-SAP simulations leaving out a block of $B = 10$ future values (see Figure 6) were overall similar to the corresponding 1-SAP simulations. In particular, PSIS-LFO-CV turned out to be close to unbiased in the approximation of exact LFO-CV. However, the accuracy of PSIS-LFO-CV for block-4-SAP turned out to be highly variable when applied to autogressive models (see the last three rows in Figure 6), something that is also visible in block-1-SAP although to a smaller degree. This seems to be a counter-intuitive result given that predictions should be more certain in the block version as more observations are available to inform the model. However, it can be explained as follows. In autoregressive models, predictions of future observations directly depend on past observations, that is predictions are not conditionally independent. This becomes a problem when dealing with observations that are missing in the approximating model right after the block of left out observations, since the directly preceeding observations are part of the block and are thus have to be treated as missing values (for details see Section 2.2). This implies a disproportionally high variability in the predictions of observations right after the block in autoregressive models, which then naturally propagates into higher variability of the PSIS-LFO-CV approximations.

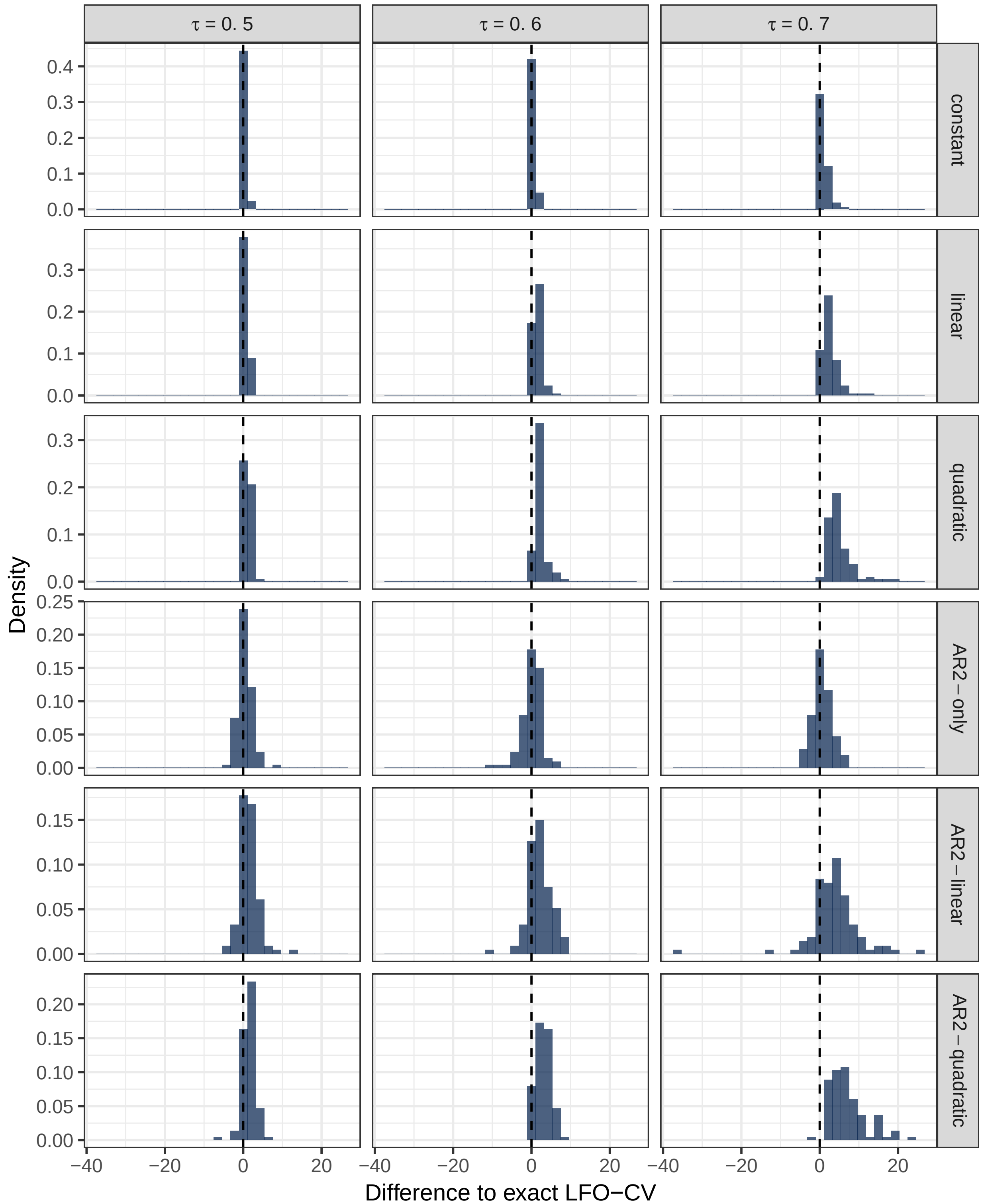

Figure 4: Simulation results of 4-step-ahead predictions.

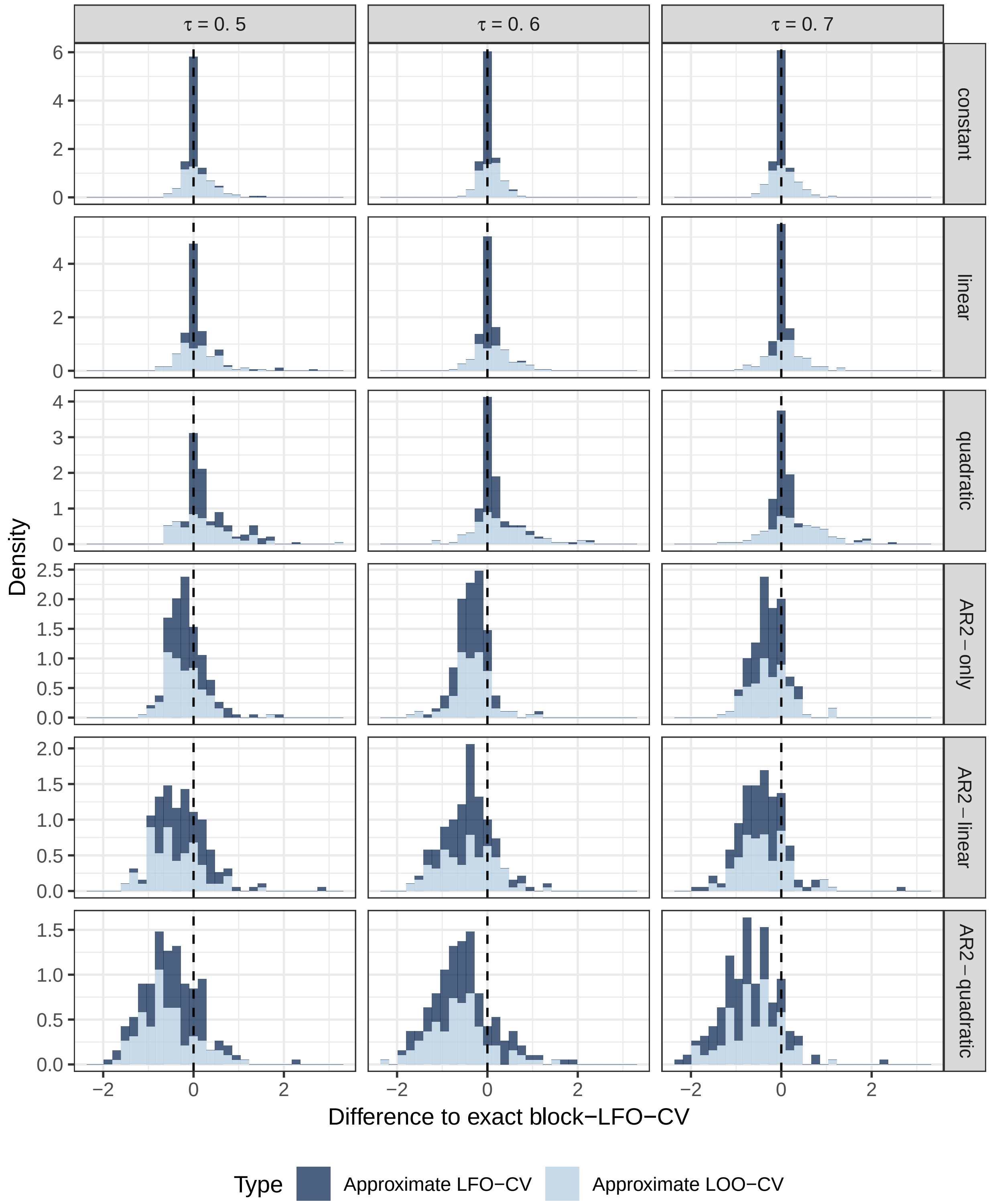

Figure 5: Simulation results of block 1-step-ahead predictions.

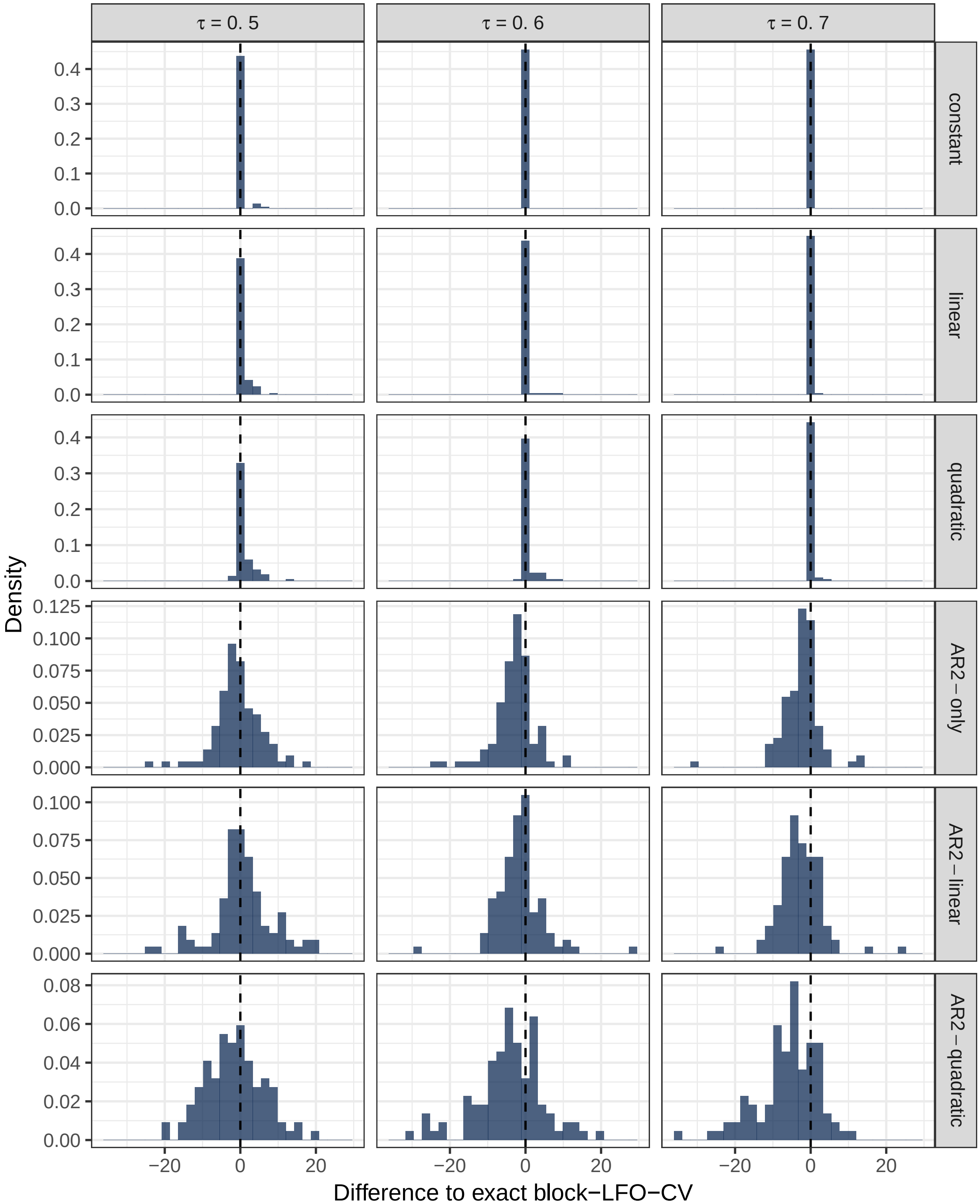

Figure 6: Simulation results of block 4-step-ahead predictions.

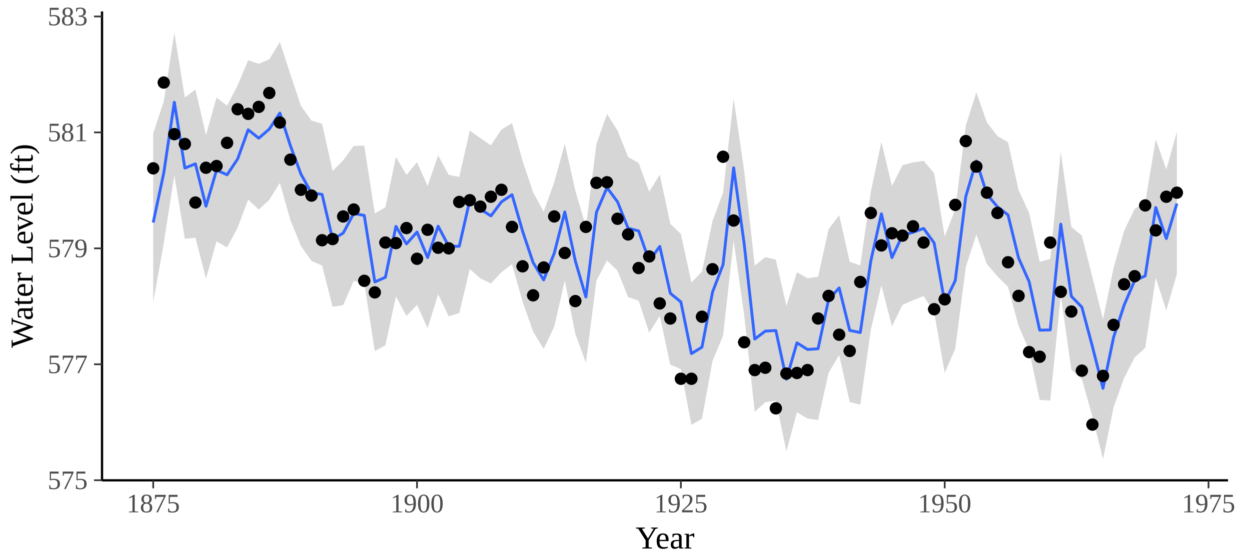

Figure 7: Water Level in Lake Huron (1875-1972). Black points are observed data. The blue line represents mean predictions of an AR(4) model with 90% prediction intervals shown in gray.

4 Case Studies

4.1 Annual measurements of the level of Lake Huron

To illustrate the application of PSIS-LFO-CV for estimating expected M-SAP performance, we will fit a model for 98 annual measurements of the water level (in feet) of Lake Huron from the years 1875–1972. This data set is found in the *datasets* R package, which is installed automatically with R (R Core Team, 2018). The time-series shows rather strong autocorrelation of the well as some trend towards lower levels for later points in time. We fit an AR(4) model and display the model implied predictions along with the observed values in Figure 7.

Based on this data and model, we will illustrate the use of PSIS-LFO-CV to provide estimates of 1-SAP and 4-SAP leaving out all future values as well as leaving out only a block of future values. To allow for reasonable predictions of future values, we will require at least $L = 20$ historical observations (20 years) to make predictions. Further, we set a threshold of $\tau = 0.6$ for the Pareto k value at which define that refitting becomes necessary. Our fully reproducible analysis of this case study can be found at (add link).

We start by computing exact and PSIS-approximated LFO-CV of 1-SAP where we leave out all future values. We compute $\text{ELPD}_{\text{exact}}$ = -93.38 and $\text{ELPD}_{\text{approx}}$ = -91.73, which are highly similar. Plotting the Pareto k estimates reveals that the model had to be refit 4 times, out of a total of $N - L = 78$ predicted observations (see Figure 8). On average, this means one refit every 19.5 observations, which implies a drastic speed increase as compared to exact LFO-CV. Performing LFO-CV of 4-SAP leaving out all future values, we compute $\text{ELPD}_{\text{exact}}$ = -538.68 and $\text{ELPD}_{\text{approx}}$ = -539.58, which are again highly similar. Although did not see this pattern in this example, in general for increasing M, the approximation tends to become less accurate in absolute ELPD units, as the ELPD increment of each observation will be based on more and

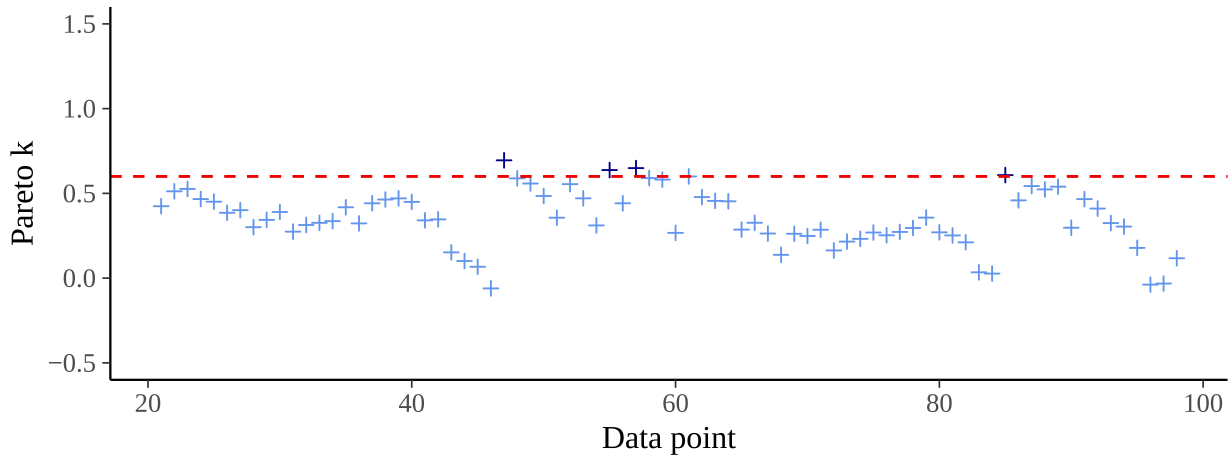

Figure 8: Pareto k estimates for PSIS-LFO-CV of the Lake Huron model leaving out all future values. The dotted red line indicates the threshold at which the refitting was necessary.

more observations. Since, for constant threshold τ, the importance weights are the same independent of M, Pareto k estimates are also the same in 4-SAP as in 1-SAP.

It is not entirely clear how stationary the time-series is as it may have a slight negative trend across time. However, the AR(4) model we are using assumes stationarity and it is appropriate to also use block-LFO-CV for this example, at least for illustrationary purposes. We choose to leave out a block of $B = 10$ future values as the dependency of an AR(4) model will not reach that far into the future. That is, we will include all observations after this block when re-fitting the model.

Approximate LFO-CV of block-1-SAP reveals $\text{ELPD}_{\text{exact}}$ = -88.55 and $\text{ELPD}_{\text{approx}}$ = -87.99, which are highly similar. Plotting the Pareto k estimates reveals that the model had to be refit 2 times, out of a total of $N - L = 78$ predicted observations (see Figure 9). On average, this means one refit every 39 observations, which again implies a drastic speed increase as compared to exact LFO-CV. What is more, we needed even fewer refits than in non-block LFO-CV, an observation we already made in our simulation in Section 3. Performing LFO-CV of block-4-SAP, we compute $\text{ELPD}_{\text{exact}}$ = -484.25 and $\text{ELPD}_{\text{approx}}$ = -488.81, which are again similar but not quite a close as in the 1-SAP case. Since AR-models fall in the class of conditionally dependent models, predicting observations right after the left-out block may be quite difficult as shown in Section 3. However, for the present data set, the PSIS apprixmations of block-LFO-CV seem to have worked out just fine.

4.2 Annual date of the cherry blossoms in Japan

The cherry blossom in Japan is a famous natural phenomenon occuring once every year during spring. As climate changes so does the annual date of the cherry blossom (Aono and Kazui, 2008; Aono and Saito, 2010). The most complete reconstruction available to date contains data between 801 AD and 2015 AD (Aono and Kazui, 2008; Aono and Saito, 2010). The data is freely available online (http://atmenv.envi.osakafu-u.ac.jp/aono/kyophenotemp4/).

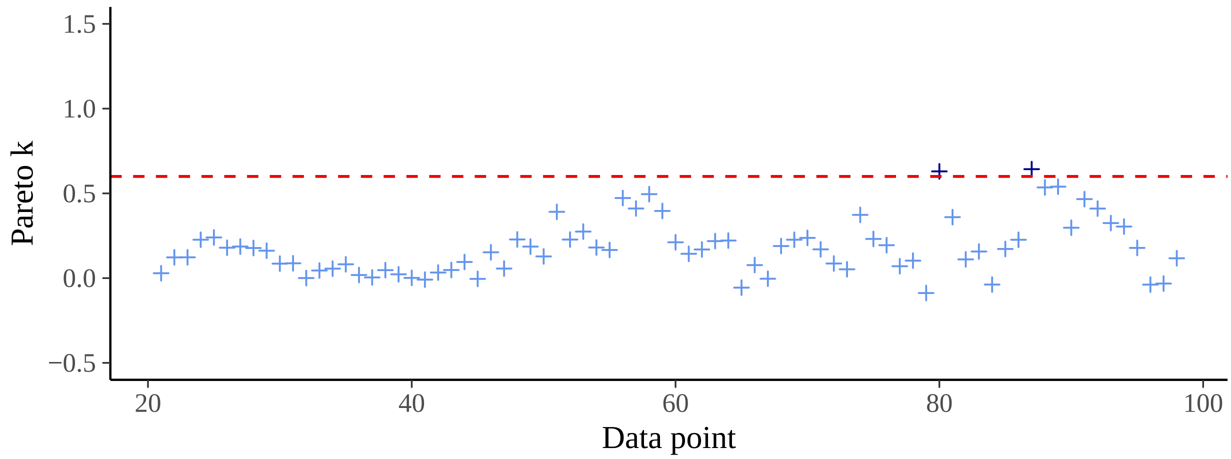

Figure 9: Pareto k estimates for PSIS-LFO-CV of the Lake Huron model leaving out a block of 10 future values. The dotted red line indicates the threshold at which the refitting was necessary.

In this case study, we are going to predict the annual date of the cherry blossom using a thin-plate regression spline (Wood, 2003) with a high number (40) of basis function to provide flexible non-linear smoothing of the time-series. A visualisation of both the data and the fitted model in provided in Figure 10. While the time-series appears rather stable across earlier centuries, with substantial variation across consequtive years, there are some clearly visible trends in the data. In particular in more recent years, the cherry blossom tended to happen much earlier than before, presumably as a result of climate change (Aono and Kazui, 2008; Aono and Saito, 2010).

Based on this data and model, we will illustrate the use of PSIS-LFO-CV to provide estimates of 1-SAP and 4-SAP leaving out all future values. To allow for reasonable predictions of future values, we will require at least $L = 100$ historical observations (100 years) to make predictions. Further, we set a threshold of $\tau = 0.6$ for the Pareto k value at which define that refitting becomes necessary. As the time series is not stationary, in particular towards the end, we do not investigate the models perfomence in terms of block-LFO-CV. Our fully reproducible analysis of this case study can be found at (add link).

We start by computing exact and PSIS-approximated LFO-CV of 1-SAP where we leave out all future values. We compute $\text{ELPD}_{\text{exact}}$ = -2349.48 and $\text{ELPD}_{\text{approx}}$ = -2346.39, which are highly similar. Plotting the Pareto k estimates reveals that the model had to be refit 39 times, out of a total of $N - L = 727$ predicted observations (see Figure 8). On average, this means one refit every 18.64 observations, which implies a drastic speed increase as compared to exact LFO-CV. Performing LFO-CV of 4-SAP leaving out all future values, we compute $\text{ELPD}_{\text{exact}}$ = -9362.47 and $\text{ELPD}_{\text{approx}}$ = -9353.94, which are again similar, but not as close as in the 1-SAP case. This coincides with our simulation results which indicated an increasing variance of PSIS-LFO-CV for increasing M. Since, for constant threshold τ, the importance weights are the same independent of M, Pareto k estimates are also the same in 4-SAP as in 1-SAP.

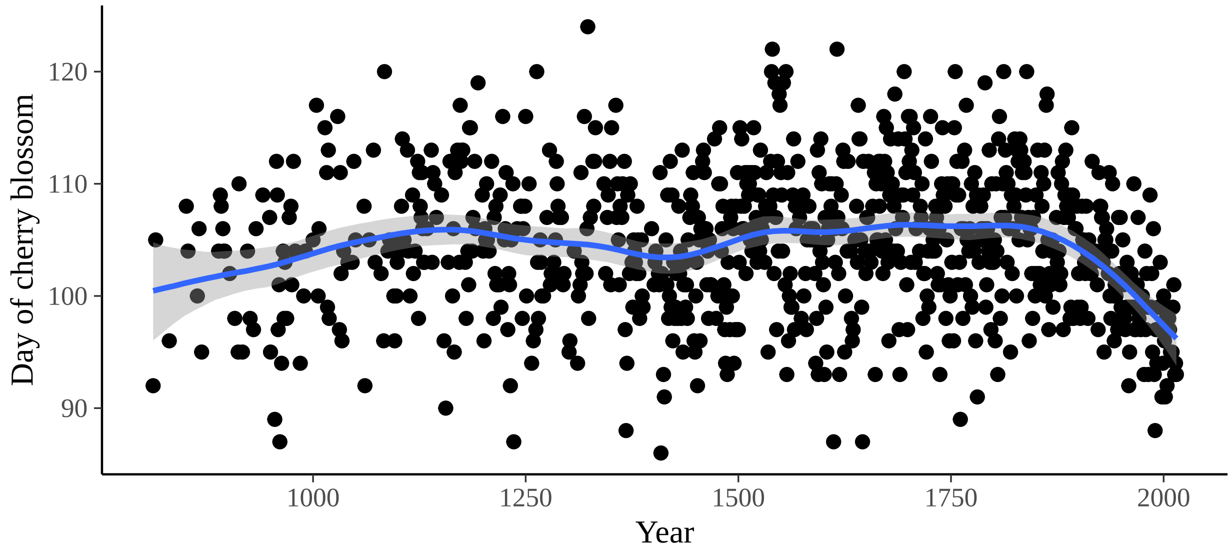

Figure 10: Day of the cherry blossom in Japan (812-2015). Black points are observed data. The blue line represents mean predictions of a thin-plate spline model with 90% regression intervals shown in gray.

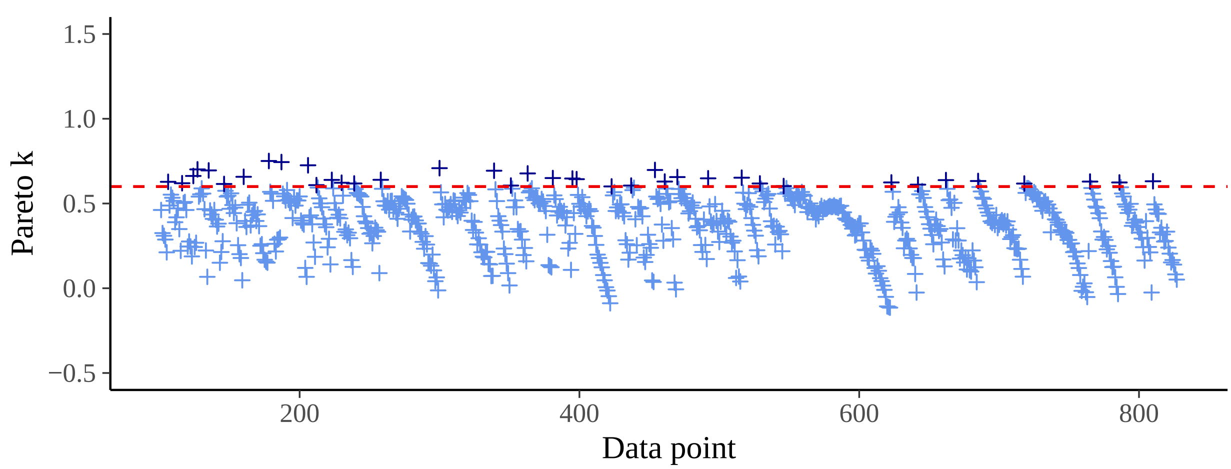

Figure 11: Pareto k estimates for PSIS-LFO-CV of the cherry blossom model leaving out all future values. The dotted red line indicates the threshold at which the refitting was necessary.

5 Discussion

In the present paper, we proposed and evaluated a new method to approximate cross-validation methods for time-series models, which we called PSIS-LFO-CV. It follows the common task of time-series models to predict future values based solely on past values. Within the set of such prediction tasks, we can choose the number M of future values to be predicted at a time and how much of the future we leave out, either all future values (M-SAP) or only a block of more recent future values (block-M-SAP).

For a set of common time-series models, we established via simulations that PSIS-LFO-CV is a close to unbiased approximation of exact LFO-CV if we choose the threshold τ of the Pareto-k-estimates to be not larger than $\tau = 0.6$. As the number of required model refits, and thus the computiational time, increases with decreasing τ, we currently see $\tau = 0.6$ as a good default when perfoming PSIS-LFO-CV. This is noticably smaller than the recommended threshold for PSIS-LOO-CV of $\tau = 0.7$, because, in PSIS-LFO-CV, observations with high Pareto-k-estimates also influence the approximation of directly preceeding observations, thus having a stronger influence on the overall accuarcy than in PSIS-LOO-CV.

Among other things, our simulations indicated that the accuracy of PSIS approximated block-M-SAP is highly variabiable for conditionally dependent models such as autoregressive models (see Section 3.1 for details). Together with the fact that block-M-SAP is only theoretically reasonable for stationary time series, as the future will always be informative for non-stationary ones, this leaves PSIS approximated block-M-SAP in a difficult spot. It appears to be a theoretically reasonable and empirically accurate choice only for conditionally independent models fit to stationary time-series. If the time-series is not too long and the corresponding model not too complex, so that a few more refits are acceptable, it might thus be more consistent and save to just using PSIS-LFO-CV of M-SAP not trying to approximate block-M-SAP at all.

Lastly, we want to briefly note that LFO-CV can also be used to compute marginal likelihoods. Using basic rules of conditional probability, we can factorize the log marginal likelihood as

$$\log p(y) = \sum_{i=1}^{N} \log p(y_i \,|\, y_{<i}). \tag{12}$$

This is nothing else than the ELPD of 1-SAP if we set $L = 0$, that is if we choose to predict **all** observations using their respective past (the very first observation is only predicted from the prior). As such, marginal likelihoods may be approximated using PSIS-LFO-CV. Although this approach is unlikely to be more efficient than methods specialized to compute marginal likelihoods such as bridge sampling (Meng and Wong, 1996; Meng and Schilling, 2002; Gronau et al., 2017), it may be a noteworthy options if, for some reason, other methods fail.

References

Ando, T. and Tsay, R. (2010). Predictive likelihood for Bayesian model selection and averaging. *International Journal of Forecasting*, 26(4):744–763.

Aono, Y. and Kazui, K. (2008). Phenological data series of cherry tree flowering in kyoto, japan, and its application to reconstruction of springtime temperatures since the 9th century. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 28(7):905–914.

Aono, Y. and Saito, S. (2010). Clarifying springtime temperature reconstructions of the medieval period by gap-filling the cherry blossom phenological data series at kyoto, japan. *International journal of biometeorology*, 54(2):211–219.

Brockwell, P. J., Davis, R. A., and Calder, M. V. (2002). *Introduction to time series and forecasting*, volume 2. Springer.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.

Bürkner, P.-C. (2018). Advanced bayesian multilevel modeling with the R package brms. *The R Journal*, pages 395–411.

Bürkner, P.-C., Gabry, J., and Vehtari, A. (in review). Leave-one-out cross-validation for non-factorizable normal models.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Ridell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., and Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of mathematical psychology*, 81:80–97.

Hamilton, J. D. (1994). *Time series analysis*, volume 2. Princeton University Press.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401.

Meng, X.-L. and Schilling, S. (2002). Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3):552–586.

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Vehtari, A., Gelman, A., and Gabry, J. (2017a). Pareto smoothed importance sampling. *arXiv preprint*.

Vehtari, A., Gelman, A., and Gabry, J. (2017b). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.

Vehtari, A. and Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural computation*, 14(10):2439–2468.

Vehtari, A. and Ojanen, J. (2012). A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228.

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.