OIKOS

Complexity is costly: comparing parametric and non-parametric methods for short-term population forecasting

Journal:	*Oikos*
Manuscript ID:	OIK-00916.R1
Wiley - Manuscript type:	Research
Date Submitted by the Author:	30-Oct-2013
Complete List of Authors:	Ward, Eric; Northwest Fisheries Science Center, National Oceanic and Atrmospheric Administration, Conservation Biology Division Holmes, Eli; Northwest Fisheries Science Center, National Oceanic and Atrmospheric Administration, Conservation Biology Division Thorson, James; Northwest Fisheries Science Center, National Oceanic and Atrmospheric Administration, Fisheries Resource and Monitoring Division Collen, Ben; University College London, Centre for Biodiversity & Environmental Research
Keywords:	time series, forecasting, density dependence

Abstract:

Short-term forecasts based on time series of counts or survey data are widely used in population biology to provide advice concerning the management, harvest and conservation of natural populations. A common approach to produce these forecasts uses time-series models, of different types, fit to time series of counts. Similar time-series models are used in many other disciplines, however relative to the data available in these other disciplines, population data are often unusually short and noisy and models that perform well for data from other disciplines may not be appropriate for population data. In order to study the performance of time-series forecasting models for natural animal population data, we assembled 2379 time series of vertebrate population indices from actual surveys. Our data were comprised of three vastly different types: highly variable (marine fish productivity), strongly cyclic (adult salmon counts), and small variance but long-memory (bird and mammal counts). We tested the predictive performance of 49 different forecasting models grouped into three broad classes: autoregressive time-series models, non-linear regression-type models and non-parametric time-series models. Low-dimensional parametric autoregressive models gave the most accurate forecasts across a wide range of taxa; the most accurate model was one that simply treated the most recent observation as the forecast. More complex parametric and non-parametric models performed worse, except when applied to highly cyclic species. Across taxa, certain life history characteristics were correlated with lower forecast error; specifically, we found that better forecasts were correlated with attributes of slow growing species: large maximum age and size for fishes and high trophic level for birds.

Complexity is costly: a meta-analysis of parametric and non-parametric methods for short-term population forecasting

Eric J. Ward[1*], Eli E. Holmes[1], James T. Thorson[1], Ben Collen[2]

1. NOAA Fisheries, Northwest Fisheries Science Center, 2725 Montlake Blvd E, Seattle, WA 98112

2. Institute of Zoology, Zoological Society of London, Regent's Park, London, United Kingdom, NW1 4RY

* Corresponding author: email: eric.ward@noaa.gov, ph: (206)-302-1745, fax: (206)-860-3217

Running head: Comparing methods for short term forecasting

Key words: Living Planet Index, RAM Legacy, Global Population Dynamics Database, machine learning, neural networks, population forecasting, non-parametric, salmon forecasting, autoregressive modeling

Abstract

Short-term forecasts based on time series of counts or survey data are widely used in population biology to provide advice concerning the management, harvest and conservation of natural populations. A common approach to produce these forecasts uses time-series models, of different types, fit to time series of counts. Similar time-series models are used in many other disciplines, however relative to the data available in these other disciplines, population data are often unusually short and noisy and models that perform well for data from other disciplines may not be appropriate for population data. In order to study the performance of time-series forecasting models for natural animal population data, we assembled 2379 time series of vertebrate population indices from actual surveys. Our data were comprised of three vastly different types: highly variable (marine fish productivity), strongly cyclic (adult salmon counts), and small variance but long-memory (bird and mammal counts). We tested the predictive performance of 49 different forecasting models grouped into three broad classes: autoregressive time-series models, non-linear regression-type models and non-parametric time-series models. Low-dimensional parametric autoregressive models gave the most accurate forecasts across a wide range of taxa; the most accurate model was one that simply treated the most recent observation as the forecast. More complex parametric and non-parametric models performed worse, except when applied to highly cyclic species. Across taxa, certain life history characteristics were correlated with lower forecast error; specifically, we found that better forecasts were correlated with attributes of slow growing species: large maximum age and size for fishes and high trophic level for birds.

Introduction

Short-term forecasts are used widely in population biology – fisheries biologists forecast commercially valuable species to inform harvest levels and to evaluate management strategies, conservation biologists use forecasts to evaluate the extinction risks for threatened species, and theoretical biologists rely on forecasts to test predictions of population responses to perturbations. The challenge, particularly with limited data, is how should predictions be made? In an infinite data universe, a mechanistic model could be constructed from first principles, incorporating population-specific biological information such as age-structured survival or fecundity rates, spatial structure or habitat information, species interactions, and sex-ratios (Hilborn & Walters 1992; Buckland *et al.* 2004; Newman *et al.* 2006). In data limited situations, however, there is little data to inform the nature of the complexity. A more common approach, taken in data-limited situations, is that population biologists apply non-mechanistic approaches to characterize patterns in the data. Types of patterns include trends, cycles, and variability. The statistical time-series models used in this non-mechanistic framework do not have a direct relationship to biological mechanisms, although they may be related to biological processes, such as population growth, survival, or density dependence.

Forecasting using this non-mechanistic approach has evolved over the last 50 years, but in population biology, the most commonly used models represent a small subset of statistical forecasting models available and used in other disciplines. To explore forecasting performance over a wide range of statistical models from the time-series modeling literature and to study which classes of models are best for the short-term prediction of population data, we adopted an inter-disciplinary approach, drawing from

63 statistical methods familiar to biologists and also approaches more frequently used in
64 other fields. We assembled a large database of natural population time series to evaluate
65 the real-world predictive accuracy of three large classes of statistical time-series models:
66 autoregressive time-series models, non-linear regression models and non-parametric
67 time-series models.

68 Autoregressive integrated moving average (ARIMA) models have a long history
69 in time-series analysis and have been widely used for population forecasting (Dennis,
70 Munholland & Scott 1991; Holmes *et al.* 2007; Ives, Abbott & Ziebarth 2010). Important
71 variants of ARIMA models include AR models, such as stochastic exponential growth
72 models and Gompertz density-dependent models, state-space models and correlated error
73 models. State-space models separate the total variance into process and observation error
74 components, yielding more precise estimates of the hidden true states of nature (e.g.
75 abundance, vital rates) when the data include high observations or error (Lindley 2003;
76 Holmes *et al.* 2007). ARIMA models with correlated errors allow the temporal deviations
77 to be temporally dependent or smoothed in different ways (Ives, Abbott & Ziebarth
78 2010). Regardless of how errors are modeled, all ARIMA models assume that the states
79 of nature at two points in time separated by a time lag p are linearly related to one
80 another. A variety of natural phenomena can lead to more complex lag structures,
81 including interactions within- and between-species (May 1977; Sugihara & May 1990),
82 age-structured demography (Gurtin & Maccamy 1974), variable sex ratios (Hassell,
83 Waage & May 1983), extrinsic forcing factors such as human disturbances, or non-linear
84 responses of species to a changing environment (Higgins *et al.* 1997; Bjornstad &
85 Grenfell 2001). The second class of models we examined, non-linear regression, provides

an approach for fitting a flexible model without specifying a linear form for the lag structure. Two types of non-linear regression models were included in this class: generalized additive models (GAMs; Wood 2006) and local regression models (e.g. 'loess'; Cleveland & Devlin 1988). The third class of models we examined, non-parametric time-series methods, treats complex lag-structure in data by allowing the lag structure to have a non-linear and non-parametric form. Several non-parametric time-series models were included in this class: projection models (Sugihara, Grenfell & May 1990; Sugihara & May 1990), neural networks (Lek *et al.* 1996), kernel regression, Gaussian process models and random forest regression (Cutler *et al.* 2007).

The properties of these parametric and non-parametric time-series methods have been studied using data from other disciplines (reviewed by Stock & Watson 1999; De Gooijer & Hyndman 2006). However, time-series data in the biological sciences present a unique set of challenges. First, population data are relatively short (typically < 25 data points; Collen *et al.* 2009) compared to the thousands of data points in financial, environmental and engineering time series. Second, population data are influenced by the presence of observation errors, resulting from uncertainty in measurement, sampling and detection rates. Unlike other fields, it is often difficult to conduct replicated survey experiments that could be used to estimate the observation error variance. As a result, the magnitude of the observation error variance is generally unknowable.

The first objective of our study was to use a meta-analysis framework to compare the short-term forecasting performance of parametric and non-parametric univariate models using our dataset of 2379 vertebrate population counts and indices. Large datasets of population time series have been used to evaluate population dynamics questions (for

example, Hilborn & Liermann 1998; Knape & de Valpine 2012) and meta-analyses of forecasting performance have been performed in other fields (Stock & Watson 1999), but to date, no large-scale forecasting meta-analysis has been carried out for ecological data, with the exception of (Stergiou & Christou 1996), who compared methods for predicting fisheries catches. However, catches may not translate well to forecasts at the population level because catches reflect a combination of population abundance, market prices, and the behavior of fishers. For similar reasons, extending meta-analysis results from other fields to ecological data is difficult because different modeling approaches perform differently for different types of data. For example, Toth, Brath & Montanari (2000) found that in predicting rainfall, neural network time-series models offered an advantage over ARIMA models, while the opposite appears to be true for macroeconomic data (Stock & Watson 1999). A further complication of previous meta-analyses is that as methods have evolved, older published studies include only a subset of the tools and models currently available.

The second objective of our analysis was to examine correlations between forecast accuracy and biological or statistical covariates (life-history characteristics, time-series length and variability). For example, our expectation was that longer time series with low levels of variation are associated with forecasts with low errors. We first explored this question on a taxonomic level and looked at whether certain classes of forecasting models work particularly well for particular taxonomic classes of organisms (birds, mammals, and fish). We then used a subset of our time series for which we had detailed biological covariates and explored whether certain attributes of species' life histories – such as growth rate, age at maturity, mean adult size or weight, trophic

position – make the abundance of these species easier to forecast. Such an analysis can guide biologists towards those forecasting models that tend to perform better for particular taxa.

Methods

Time-series data

We compiled a database of 2379 univariate time series of aquatic and terrestrial vertebrates worldwide (Table 1). Only time series with at least 25 continuous observations (no missing values) were included. Most of the time series were population counts or indices of abundance, but we also included time series of marine fish production (recruits per spawning stock biomass) in our database. We assembled bird and mammal abundance time series from the Living Planet Index (LPI) Database, the North American Breeding Bird Survey (BBS), and the Royal Society for the Protection of Birds (RSPB), salmon spawner abundance data from published literature (Holmes & Fagan 2002; Dorner, Peterman & Haeseker 2008), the National Marine Fisheries Service (Ford 2011) and StreamNet, and marine fish productivity from the RAM Legacy database (Ricard *et al.* 2011). Time series were filtered to only include those collected from a consistent survey of some type.

The LPI Database (Loh *et al.* 2005; Collen *et al.* 2009) is a database of worldwide population time series, collated from published scientific literature and other global databases, especially the Global Population Dynamics Database (NERC Centre for Population Biology 2010) and the Pan-European Common Bird Monitoring Scheme (Pan-European Common Bird Monitoring Scheme 2011). The North American BBS

(Sauer *et al.* 2011; Risely *et al.* 2012) is monitoring program by the U.S. Geological Survey's Patuxent Wildlife Research Center and Environment Canada's Canadian Wildlife Service. It provides regional population estimates from standardized roadside route surveys for North American breeding birds. The RSPB breeding bird data were compiled by the RSPB from data collected by the Statutory Conservation Agencies/RSPB annual breeding bird scheme, the Rare Breeding Birds Panel, and RSPB's own bird monitoring programs. These data consist of estimated population sizes for 61 rare or scarce breeding bird species in the United Kingdom based on censuses of known breeding sites. Our Pacific Northwest salmon data consist of yearly spawner counts of Chinook (*Oncorhynchus. tshawytscha*), pink (*O. tshawytscha*), chum (*O. keta*), coho (*O. kisutch*), and sockeye salmon (*O. nerka*) in British Columbia, Canada and Washington, Oregon, and California, USA collected as part of state and provincial monitoring programs. The RAM Legacy database includes time series of fish biomass and productivity (recruits/spawning stock biomass) for marine fishes around the globe. We only included productivity time series in our database because the RAM Legacy adult spawning biomass time series are smoothed output from stock assessment models.

Biological covariate data

To test whether certain groups of species are more predictable than others, we assembled biological covariates for species in our three largest datasets: marine fish productivity, bird counts and salmon abundance. For species in the marine fish productivity dataset, we assembled maximum age, mean adult length, relative weight, and trophic level information from RAM Legacy and FishBase (Froese & Pauly 2000).

178 Relative weight is a proxy for the girth of each species, calculated as the residuals of log
179 length-log weight regressions. Weight by itself was not included as a covariate because
180 weight and length are highly correlated. For the bird species in the BBS, RSPB and LPI
181 datasets, we assembled mean adult weight, generation length, and trophic level
182 information from the LPI database and BirdLife International. For the database of adult
183 salmon counts, we assembled mean length of spawning adults and trophic level for each
184 species from FishBase (Froese & Pauly 2000).
185
186 *Time-series models*
187 We tested the forecasting performance of 49 univariate time-series models. These
188 models can be classified into three groups: ARIMA models, regression models and non-
189 parametric models. We summarize the models below and more details, including the R
190 functions to implement each model, are available in the SI.
191 1. ARIMA models
192 ARIMA stands for autoregressive integrated moving average and is a model that
193 combines autoregressive (AR), differencing (I), and moving average (MA) components.
194 An AR model of logged-abundance (Y_t) takes the form

$$Y_t = b_1 Y_{t-1} + b_2 Y_{t-2} + \ldots + b_q Y_{t-p} + e_t$$

195 A MA model is similar but instead of Y being autoregressive, the error term (e_t) is
196 modeled as autoregressive. A model that combines both AR and MA components is
197 ARMA, and if the differences (Y_t-Y_{t-1}, Y_t-Y_{t-2}, etc.), rather than Y, are treated as the
198 response, the result is an ARIMA model. All of these models can be written in
199 ARIMA(p, d, q) form in terms of three parameters: p, the number of autoregressive

terms, *d*, the degree of differencing, and *q*, the number of moving average terms. See Ives, Abbott & Ziebarth (2010) for a discussion of ARIMA models used in ecology and the SI for more details.

The most basic ARIMA model we considered was a random walk model, denoted ARIMA(p = 0, d = 1, q = 0), with and without drift. We also considered state-space versions of these models (Holmes 2001; Lindley 2003; Holmes *et al.* 2007), which include an observation model in addition to the process model. Potentially unrealistic assumptions made by the simple random walk are that (1) the mean trend is constant through time, (2) stochastic fluctuations through time are independent and temporally uncorrelated, and (3) that population change is not density-dependent. To relax assumptions (2) and (3), we fit a range of different ARIMA models to include temporally correlated errors and mean-reversion (density-dependence). Random walks with density-dependence (Gompertz random walks; Dennis *et al.* 2006), are ARIMA(1,0,0) with a constant, random walks with autocorrelated errors are ARIMA(1,1,0), random walks with smoothed errors (MA) are ARIMA(1,0,1), and exponentially smoothed time series (Hyndman *et al.* 2002) are ARIMA(0,1,1). We fit a range of ARIMA models, varying *p*, *d*, and *q* from 0 to 2. All models are listed in Table 2 in the SI. Finally to relax assumption (1), we fit stochastic level models with the random walk drift parameter itself modeled as a random walk.

2. Linear and non-linear regression

We explored three types of parametric regression methods. The first was simple linear regression of logged abundance or productivity against time with temporally uncorrelated errors. Using a moving average model, ARIMA(0,0,1), we also fit a linear

223 regression with autocorrelated errors. Second we fit local regression models (Cleveland
224 & Devlin 1988), which fit local polynomial models to a specified number of neighboring
225 data points. Lastly, we evaluated non-linear regression using GAMs (Wood 2006) with
226 the degree of smoothness selected by cross validation. GAMs model the expected value
227 of a data point as a function of a link function and splines, whereas local regression uses a
228 moving window approach to sequentially fit polynomial splines to batches of data. All
229 parametric models were fit with Gaussian errors to log transformed data.

230 3. Non-parametric methods

231 We tested a variety of non-parametric methods: kernel regression, neural
232 networks, Gaussian process models, projection models and random forest regression.
233 Non-parametric kernel regression models use a kernel function to weight the importance
234 of neighboring points. Neural network time-series methods (Toth, Brath & Montanari
235 2000; Thrush, Coco & Hewitt 2008) estimate 'hidden layers' as the sum of logistic-
236 transformed inputs to relate historical observations to future states (we considered up to 3
237 hidden layers). Gaussian process models estimate the covariance between pairs of
238 neighboring observations but do not impose a parametric form for the errors nor a
239 specific lag structure. A related non-parametric approach is projection methods (S-MAP
240 and Simplex projection) which map the response value Y_t as a function of lagged
241 abundances, Y_{t-1}, Y_{t-2},…. S-MAP (Sugihara 1994) and Simplex projection (Sugihara,
242 Grenfell & May 1990) have been successful at forecasting non-linear ecological time
243 series (Hsieh, Anderson & Sugihara 2008; Glaser *et al.* 2011). Simplex uses only a few
244 neighboring points to make predictions, while S-MAP uses a distance-weighting method.
245 We implemented both approaches while automatically selecting the lagging dimensions

for each. As a final method, we tested random forest regression (Cutler *et al.* 2007), which uses lagged abundances as the predictors and uses decision trees to optimize the predictive ability. Lagged abundances at 1 to 5 time steps were used as predictors and automatically selected from decision trees with up to 5 nodes.

Model fitting and projection

Each time series was log-transformed to achieve approximate normality and to account for population growth being a multiplicative process. Time series were detrended as part of the fitting process for stationary ARIMA models (but the trend was included in model forecasts). The models were fit to the entire time series minus the last 5 time steps; this is the 'training' data. The last 5 time steps were held out to gauge predictive performance. All models were fit in R using add-on packages (R Core Development Team 2010); code and functions are provided in the SI. From the fitted models, we forecasted the next 1 to 5 years using the prediction functions supplied with the corresponding R packages (or our own function for S-MAP and Simplex projection).

Evaluation of forecast performance

Though forecast performance can be improved in some situations with ensemble forecasting from multiple models (Newbold & Granger 1974; Raftery *et al.* 2005) or by combining information across time series (Hsieh, Anderson & Sugihara 2008; Ward *et al.* 2010), our goals were to evaluate the performance of individual models and to identify which models (or model classes) are best on average across large datasets, following the approach of (Geweke, Meese & Dent 1983). Model performance in prediction (or

explanation) can be viewed through the lens of the bias-variance tradeoff, Error = Variance + Bias^2 + Irreducible error, where bias decreases and variance increases with model complexity, and irreducible error represents the unexplained variation (Burnham and Anderson 2002). When comparing the performance of multiple models across multiple time series from diverse environments and taxa, scale invariant metrics need to be used because different time series have different scales of variation. Thus, scale-dependent metrics like root mean square error (RMSE) should not be used (Hyndman & Koehler 2006). A variety of scale-invariant measures of forecasting accuracy exist. We used the mean absolute scaled error (MASE) recommended by (Hyndman & Koehler 2006). MASE allows comparison of predictive accuracy across datasets with different scales of variation and is less sensitive to extreme values and outliers.

For a single time series, the absolute scaled error (ASE) for a prediction $\hat{Y}_t$ at time t after the training data (the portion of the time-series used for fitting) is

$$ASE_t = \frac{|Y_t - \hat{Y}_t|}{\frac{1}{n-1}\sum_{i=2}^{n}|Y_i - Y_{i-1}|}$$

where Y_t is the observed value at time-step t (1 to 5) after the end of the training data (Hyndman & Koehler 2006). ASE values are calculated independently for each forecasting model. The absolute error is scaled by the mean absolute error within the training data, $\frac{1}{n-1}\sum_{i=2}^{n}|Y_i - Y_{i-1}|$, where Y_i is the i-th observation within the training data and n is the number of training observations. To calculate MASE$_t$ for a given model the ASE$_t$ values from all time series are averaged. A general property of MASE is that as time-series length increases, forecasts using a random walk without drift will converge to a MASE of 1. For short time series, such as those used here, the same random walk

model will produce MASE values higher than 1, because the small-sample mean absolute error (the denominator in the ASE equation) is an estimate of the large-*n* mean absolute error. Thus, with short time series, we compare MASE values to the MASE from the random walk without drift model (termed 'RW-MASE'). This will be some value greater than 1 for short time series. When a model has a MASE less than RW-MASE, it indicates that (1) there is structure in the data beyond that implied by a single random-walk process and (2) the model successfully models that structure to give a better forecast. MASE values higher than RW-MASE indicate that the model is either over-fitting the data or fitting an improper model to the data.

We computed MASE for 1- to 5-step ahead predictions. For each model and each time series, we predicted the future values of the times series at t=1 to 5 past the end of the training data, giving us $\hat{Y}_1, \ldots, \hat{Y}_5$. With these and the observed values, $Y_1, \ldots, Y_5$, we computed the ASE and MASE statistics for each model.

Identifying covariates useful in prediction

We conducted a secondary analysis to explore which statistical and biological covariates were correlated with better predictive accuracy (lower ASE values). For this analysis, we used only time series for species with covariate information: birds (n=890) from the BBS, RSPB and LPI datasets, marine fish (n=133) from the RAM Legacy productivity dataset, and salmon (n=289) from our combined salmon dataset. In addition to biological covariates, we included the following descriptive statistics as covariates: time-series length, variance of the lag-1 differences, lag-1 autocorrelation (calculated as the ACF of differenced observations), mean trend, current abundance relative to the

maximum observed (a measure of depletion), and the ratio of observation to process variance as estimated by a state-space random walk with drift model.

For the response variable, we used the natural log of the average ASE statistic from the GAM model for forecasts 1 to 3 time steps ahead:

$$\overline{\text{ASE}} = \frac{\sum_{t=n+1}^{n+3} |Y_t - \hat{Y}_t| / 3}{\frac{1}{n-1}\sum_{i=2}^{n} |Y_i - Y_{i-1}|}$$

Here, $\hat{Y}_t$ is the estimate for time t from the GAM model fit to a single time series and Y_t is the actual observed value at time t. ASE values 1 to 3 time steps ahead were averaged because using an ASE value for one time step alone is highly sensitive to outliers. Using $\overline{\text{ASE}}$ reduced the effect of outlier values. We show the results using the $\overline{\text{ASE}}$ values using $\hat{Y}_t$ from the GAM model, however we did the analysis with $\overline{\text{ASE}}$ computed with $\hat{Y}_t$ values from the ARIMA models, and results were similar. Separate linear regressions of covariates against $\overline{\text{ASE}}$ were used for the bird, marine fish productivity, and salmon time series to prevent results from being dominated by the taxa with greater sample size. Stepwise regression with AIC as a model selection tool was used to identify covariates with higher explanatory power.

Results

We summarized the forecast accuracy of different classes of models using the mean absolute scaled error (MASE) statistic (Hyndman & Koehler 2006). This metric allows forecast accuracy for different datasets to be compared on a similar scale and combined into a single number, thus allowing us to evaluate forecast performance integrated over multiple time series. Examining MASE across taxonomic groups (birds,

marine fish productivity, salmon counts, mammal abundance), we found that GAMs and low dimensional ARIMA models (of various types including AR and ARMA, but excluding pure MA models) produced short-term forecasts with the best predictive accuracy. No particular ARIMA model stood out; rather, the well-performing ARIMA models were characterized by simplicity (few estimated parameters) and a strong connection between the forecast and the last observed value. The worst performing methods included linear regression, neural network models, S-MAP projection and local regression (Fig. 1). Although GAM and simple ARIMA models performed best, their MASE statistics were similar to that of a random walk without drift (the baseline model) for birds, mammals, and marine fish productivity, and their predictions became steadily worse for 2, 3, and 4 time steps forward (Fig. 1). ARIMA models only outperformed the baseline random walk when applied to data from highly cyclic salmon species. For some salmon species, 2- and 4-step ahead forecasts were just as good as 1-step ahead forecasts (Fig. 2). These results were particularly true for pink and sockeye salmon – species whose life histories cause regular population cycles with even-numbered periods. For these two cyclic species, some non-parametric methods (e.g. Simplex projection and random forest regression) did as well as the ARIMA models (Fig. 2), presumably because they capture the lagged structure in the time series. While the ARIMA models in Fig. 1 do not include lags greater than 1, they are able to model lag-2 cycles via negative autocorrelation between t and t-1. Detailed results for all models are given in Table S2 in the SI.

Results from our analysis of covariates and forecasting performance identified biological and statistical covariates associated with better forecasts (lower errors),

however the covariates selected depended on the taxa. For the marine fish productivity dataset, we found that species with larger maximum lengths and larger maximum ages were associated with improved forecasts (Table 2). In terms of the biological effect size, we found the effects of length and maximum age to be equivalent (Fig. 3). We also found that an increasing ratio of observation to process variance was correlated with lower forecast error – meaning that when observation variance contributed a larger proportion of the total variance, the relative influence of process variance was smaller, and the forecasts tended to have lower error (relative to the variance in the time series). For the bird dataset, the only biological variable associated with better forecasts was trophic level; the positive relationship indicates that higher trophic level species in our dataset were associated with lower forecast errors. Two statistical covariates were also associated with better forecasts for birds: decreased total variance in the time series and increased autocorrelation (Table 2). No significant biological or statistical predictors were found for the combined salmon datasets, possibly because the small number of species included (five) provided low resolution. Although these results are for forecasts from the GAM model, we found similar covariates when we used forecasts from the ARIMA models. This is not surprising since the forecasts (and ASE or MASE values) from the GAMs and ARIMA models are correlated.

Discussion and Conclusions

Historically, the majority of ecological time series analysis has focused on identifying explanatory processes (competition, density dependence, Allee effects). These model selection analyses have used statistics such as Type I error rates, or model

selection tools like AIC to identify models that balance the explanatory ability of models with predictive ability (this is the principle the parsimony; Burnham and Anderson 2002). Less work has been done to investigate the predictive or forecasting ability of statistical models in ecology. Short-term forecasts are becoming widely used in population biology, and in this paper, we sought to identify specific classes of models that (1) are flexible enough to fit a range of population processes, from declines to density dependence, and (2) have low prediction error. These characteristics are particularly important for species at risk, or species that are commercially valuable (such as fish populations). In data-rich situations, population forecasts might be improved by including biological mechanisms and dynamics (though including mechanisms may also yield worse fits; Perretti et al. 2013). In data-poor situations, a time series of estimates of abundance or biomass is often the only information available. An ever-increasing array of modeling approaches can be used to make short-term forecasts using only time-series data and have been used in other disciplines, however the performance of these approaches may be quite different for animal population data given its typically noisy and short nature. Our meta-analysis of vertebrate time series included species from aquatic and terrestrial ecosystems and diverse data types: we included highly variable data (marine fish), low variability data (birds, mammals), data with cyclic dynamics (salmon counts), and data across a gradient of species longevity.

For forecasting species without strong cyclic dynamics (birds, mammals, marine fish), we found the best performers to be GAMs and ARIMA models, which includes random walks with drift, models with temporally correlated or smoothed errors, state-space models, and ARIMA models with a lag-1 correlation. However, averaged over all

non-cyclic species, both small and short-lived and large and long-lived, the 'best' models for these non-cyclic species only did as well or slightly better than a random walk without drift (Fig. 1; Table S2 in SI). Effectively, this means that the forecast involving the fewest estimated parameters, which effectively simply uses the last observation at time t, was the best prediction of the value of the population at time $t+k$ (k=1:5). This highlights the cost of trying to estimate even the trend (drift), much less more complex lag structure, when using short, noisy time series with unknown levels of observation error. That these models did not strongly outperform the baseline random walk without drift was surprising since time series from all taxa in our analysis showed evidence for a lag-1 negative autocorrelation (Fig. 4). Such negative autocorrelation is common in population data and can be generated by age-structured demography (especially for semelparous species, such as salmon), sex-ratios, density-dependence, and observation errors. However for short time series, we found that estimation of these lag terms is very costly, much like Ives, Abbott & Ziebarth (2010) found, and that estimation of the observation error variance also comes at a high cost, an issue also discussed by Holmes *et al.* (2007). In the context of bias-variance tradeoff, these more complex models might fit a training dataset well, but will have low predictive power when applied to out of sample data (Burnham and Anderson 2002).

The other models types, other than ARIMA and GAMs, however, did considerably worse than baseline random walk without drift (and worse that ARIMA and GAM models). Linear regression and neural network models did especially poorly, likely due to the fact that their forecasts are not tied directly to the last observation. S-MAP, Simplex and random forest regression also did poorly for birds, mammals and marine

fish, possibly because these methods are more data intensive as they involve sampling from the lag-*p* differences in the data and thus may be especially affected by low sample size.

For the salmon time series, in contrast, we found that all ARIMA models outperformed the baseline random walk without drift. Time series of adult salmon abundance are often characterized by strong and regular cyclic patterns, producing negative correlation in the lag-1 errors. When we looked at the individual salmon species, we saw that the better performance of the ARIMA models was driven mainly by better performance for pink, sockeye, and chum salmon. Though patterns vary regionally, these three species are characterized by regular cyclic behavior (Ruggerone *et al.* 2010). GAMs, neural networks, Simplex and random forest models also did especially well for these cyclic species, though these same models performed worse than the baseline random walk when applied to less cyclic salmon species. The unusually good performance of neural networks, Simplex and random forest models for species with strong cycles highlights the ability of these non-parametric approaches to model complex structure in data.

Most of the results from our analysis of biological covariates associated with better prediction match intuition; across taxa, bird and mammal population abundance was generally forecasted with better accuracy than fish abundance or productivity (Fig. 3), and within taxa, species that are larger, older, or occupy higher trophic levels are generally easier to predict than smaller, fast growing species (Table 2). Smaller species, such as sardine or anchovies in our data, are conventionally associated with more r-selected life history types and more eruptive population dynamics. The average 1- to 3-

449 step ahead ASE statistics were larger for these species, suggesting that a random walk
450 with no drift would provide as good of a forecast as any more complicated model.
451 However, for species that were larger, were at a higher trophic level, or had larger
452 maximum ages, use of a GAM or any of the low-dimensional ARIMA models improved
453 forecasts. This suggests that low-dimensional models could also provide better than
454 random-walk forecasts for the non-cyclic species but in general only for the subset of
455 these species with larger size and higher trophic level.

456 The baseline model used in our analysis was a simple random walk without drift.
457 For this model, the *t*-step ahead forecast is simply the last observed value. No additional
458 model parameters are estimated for the actual forecast, though the calculation of the ASE
459 (the prediction error) uses an estimate of the total variance (as do all models). The failure
460 of the more complicated time-series models to provide short-term predictions with lower
461 error than the random walk without drift emphasizes 1) the cost of estimating parameters
462 in the face of noise and 2) the cost of basing short-term predictions on parameters, like
463 the trend over the whole time series, which may be more associated with long-term
464 dynamics rather than short-term behavior. For short population time series, we can
465 recommend the use of more complex forecasting models only when time series have
466 strong internal structure (e.g. the cyclic dynamics in salmon) or have lower variability
467 and higher temporal autocorrelation (larger species with higher maximum ages or higher
468 trophic level). In summary, fitting models with many parameters and the flexibility to
469 model complex structure may be tempting, but this involves estimating structure from
470 few data points. We found that estimation of even one or two parameters imposes a high

cost with little benefit for *short-term* forecasts of population abundance for species without obvious cyclic population dynamics.

Acknowledgements

We are extremely grateful for all of the hard work by the many researchers who assembled, checked, or continue to maintain the databases used in our analysis. We also thank the individuals that have created and shared R libraries and packages for time series analysis with the scientific community on CRAN, as well as Ethan Deyle, Hao Ye, and Sarah Glaser for help in implementing and testing S-Maps and Simplex. The RAM Legacy database has dozens of contributors, including (but not limited to) Julia Baum, Olaf Jensen, Coilin Minto, Ram Myers, and Kate Stanton. The RSPB bird census data was provided by Richard Gregory (The Royal Society for the Protection of Birds). The North American BBS data was provided by John Sauer (USGS Patuxent Wildlife Research Center). The bird metadata used in this study were provided by BirdLife International and were assembled as part of the Red-Flags and Extinction Risk' Working Group supported by the National Center for Ecological Analysis and Synthesis (Santa Barbara, CA, USA). We thank especially Red-Flag members Stuart Butchart, Marta Nammack, Resit Akcakaya, and David Keith who provided and assembled time series and biological covariate metadata. We thank Randall Peterman and John Froeschke for providing helpful reviews of an early draft of the manuscript.

References

Bjornstad, O.N. and Grenfell, B.T. 2001. Noisy clockwork: time series analysis of population fluctuations in animals. – Science 293: 638-643.

Buckland, S.T. et al. 2004. State-space models for the dynamics of wild animal populations. – Ecol. Mod. 171: 157-175.

Burnham, K.P. and Anderson, D.R. 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. - Springer.

Cleveland, W.S. and Devlin, S.J. 1988. Locally weighted regression: an approach to regression analysis by local fitting. – J. Am. Stat. Assoc. 83: 596-610.

Collen, B. et al. 2009. Monitoring change in vertebrate abundance: the Living Planet Index. – Cons. Bio. 23: 317-327.

Cutler, D.R. et al. 2007. Random forests for classification in ecology. – Ecology 88: 2783-2792.

De Gooijer, J.G. and Hyndman, R.J. 2006. 25 years of time series forecasting. – Int. J. Forecasting 22: 443 – 473.

Dennis, B. et al. 1991. Estimation of growth and extinction parameters for endangered species. – Ecol. Monogr. 61: 115-143.

Dennis, B. et al. 2006. Estimating density dependence, process noise, and observation error. – Ecol. Monogr. 76: 323-341.

Dorner, B. et al. 2008. Historical trends in productivity of 120 Pacific pink, chum, and sockeye salmon stocks reconstructed by using a Kalman filter. – Can. J. Fish. Aquat. Sci. 65**:** 1842-1866.

514 Ford, M.J. (ed.) 2011. Status review update for Pacific salmon and steelhead listed
515 under the Endangered Species Act. U.S. Department of Commerce, NOAA
516 Technical Memorandum, NMFS-NWFSC-113. Seattle, WA.
517 Froese, R. and Pauly, D. 2000. FishBase 2000: concepts, design and data sources.
518 ICLARM, Los Baños, Laguna, Philippines.
519 Geweke, J. et al. 1983. Comparing alternative tests of causality in temporal systems:
520 analytic results and experimental evidence. – J. Econometrics 21: 161-194.
521 Glaser, S.M. et al. 2011. Detecting and forecasting complex nonlinear dynamics in
522 spatially structured catch-per-unit-effort time series for North Pacific
523 albacore (Thunnus alalunga). – Can. J. Fish. Aquat. Sci. 68: 400-412.
524 Gurtin, M.E. and Maccamy, R.C. 1974. Non-linear age-dependent population
525 dynamics. – Arch. Ration. Mech. An. 54: 281-300.
526 Hassell, M.P. et al. 1983. Variable parasitoid sex-ratios and their effect on host-
527 parasitoid dynamics. – J. Anim. Ecol. 52: 889-904.
528 Higgins, K. et al. 1997. Stochastic dynamics and deterministic skeletons: population
529 behavior of Dungeness crab. – Science 276: 1431-1435.
530 Hilborn, R. and Liermann, M. 1998. Standing on the shoulders of giants: learning
531 from experience in fisheries. – Rev. Fish Biol. Fisher. 8: 273-283.
532 Hilborn, R. and Walters, C.J. 1992. Quantitative Fisheries Stock Assessment: Choice,
533 Dynamics, and Uncertainty. Kluwer Academic Publishers, London.
534 Holmes, E.E. 2001. Estimating risks in declining populations with poor data. – P Natl
535 Acad Sci USA 98: 5072-5077.

536 Holmes, E.E. and Fagan, W.E. 2002. Validating population viability analysis for
537 corrupted data sets. – Ecology 83: 2379-2386.
538 Holmes, E.E. et al. 2007. A statistical approach to quasi-extinction forecasting. – Ecol.
539 Lett. 10: 1182-1198.
540 Hsieh, C.H. et al. 2008. Extending nonlinear analysis to short ecological time series. –
541 Am. Nat. 171: 71-80.
542 Hyndman, R.J. and Koehler, A.B. 2006. Another look at measures of forecast
543 accuracy. – Int. J. Forecasting 22: 679-688.
544 Hyndman, R.J. et al. (2002) A state space framework for automatic forecasting using
545 exponential smoothing methods. – Int. J. Forecasting 18: 439-454.
546 Ives, A.R. et al. 2010. Analysis of ecological time series with ARMA(p,q) models. –
547 Ecology 91: 858-871.
548 Knape, J. and de Valpine, P. 2012. Are patterns of density dependence in the Global
549 Population Dynamics Database driven by uncertainty about population
550 abundance? – Ecol. Lett. 15: 17-23.
551 Lek, S. et al. 1996. Application of neural networks to modelling nonlinear
552 relationships in ecology. – Ecol. Model. 90: 39-52.
553 Lindley, S.T. 2003. Estimation of population growth and extinction parameters from
554 noisy data. – Ecol. Appl. 13: 806-813.
555 Loh, J. et al. 2005. The Living Planet Index: using species population time series to
556 track trends in biodiversity. – Philos. T. Roy. Soc. B 360: 289-295.
557 May, R.M. 1977. Thresholds and breakpoints in ecosystems with a multiplicity of
558 stable states. – Nature 269: 471-477.

559 NERC Centre for Population Biology 2010. The Global Population Dynamics
560 Database Version 2. Imperial College.
561 Newbold, P. and Granger, C.W.J. 1974. Experience with forecasting univariate time
562 series and combination of forecasts. – J. Roy. Stat. Soc. A 137: 131-165.
563 Newman, K.B. et al. 2006. Hidden process models for animal population dynamics. –
564 Ecol. Appl. 16: 74–86.
565 Pan-European Common Bird Monitoring Scheme (2011) European common bird
566 index: population trends of European common birds 2011 update. (ed. E.B.C.
567 Council). Prague.
568 Perretti C.T. et al. 2013. Model-free forecasting outperforms the correct mechanistic
569 model for simulated and experimental data. –Proceedings of the National Academy of
570 Sciences USA 110:5253–5257.
571 R Core Development Team (2010) R: A language and environment for statistical
572 computing. R Foundation for Statistical Computing , Vienna, Austria. URL =
573 http://www.R-project.org.
574 Raftery, A.E. et al. 2005. Using Bayesian model averaging to calibrate forecast
575 ensembles. – Mon. Weather Rev. 133: 1155-1174.
576 Ricard, D. et al. 2011. Examining the knowledge base and status of commercially
577 exploited marine species with the RAM Legacy Stock Assessment Database. –
578 Fish and Fisheries, 13: 380–398.
579 Risely, K. et al. 2012. The Breeding Bird Survey 2011. BTO Research Report 624.
580 Thetford.

581 Ruggerone, G.T. et al. 2010. Magnitude and trends in abundance of hatchery and
582 wild pink salmon, chum salmon, and sockeye salmon in the North Pacific
583 Ocean. – Mar. Coast. Fish. 2: 306-328.
584 Sauer, J.R. et al. 2011. The North American Breeding Bird Survey, results and
585 analysis 1966 - 2010 version 12.07.2011. (ed. U.P.W.R. Center). Laurel, MD.
586 Stergiou, K.I. and Christou, E.D. 1996. Modelling and forecasting annual fisheries
587 catches: comparison of regression, univariate and multivariate time series
588 methods. – Fish. Res. 25: 105-138.
589 Stock, J.H. and Watson, M.W. 1999. A comparison of linear and nonlinear univariate
590 models for forecasting macroeconomic time series, in R.F. Engle and H. White
591 (eds.), Cointegration, Causality, and Forecasting: A Festschrift in Honor of
592 Clive W.J. Granger, Oxford University Press, Oxford.
593 Sugihara, G. 1994. Nonlinear forecasting for the classification of natural time-series.
594 – Philos. T. Roy. Soc. A 348: 477-495.
595 Sugihara, G. et al. 1990. Distinguishing error from chaos in ecological time series. –
596 Philos. T. Roy. Soc. B 330: 235-251.
597 Sugihara, G. and May, R.M. 1990. Nonlinear forecasting as a way of distinguishing
598 chaos from measurement error in time series. – Nature 344: 734-741.
599 Thrush, S.F. et al. 2008. Complex positive connections between functional groups
600 are revealed by neural network analysis of ecological time series. – Am. Nat.
601 171: 669-677.
602 Toth, E. et al. 2000. Comparison of short-term rainfall prediction models for real-
603 time flood forecasting. – J. Hydrol. 239: 132-147.

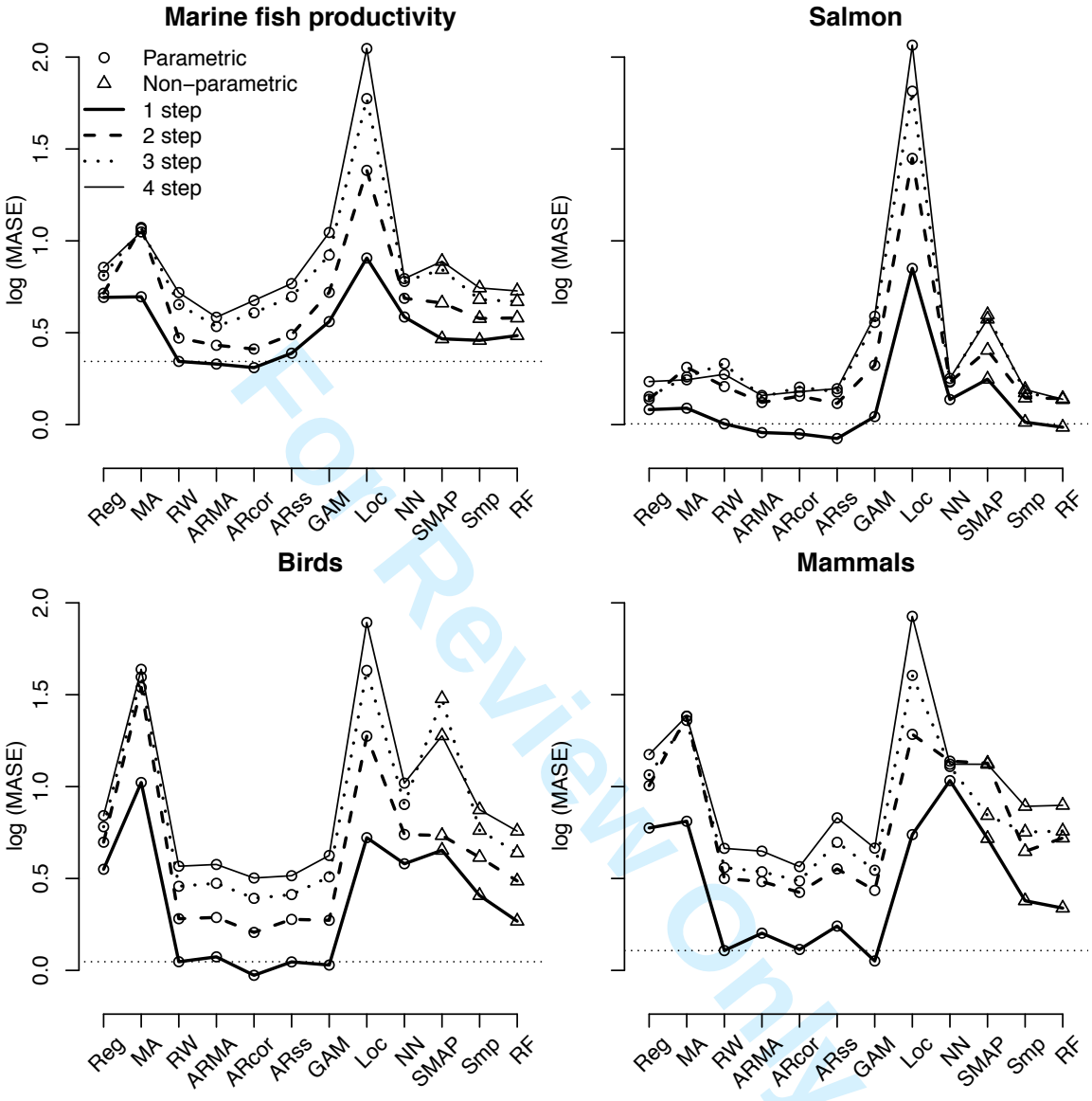
604 Ward, E.J. et al. 2010. Inferring spatial structure from time-series data: using
605 multivariate state-space models to detect metapopulation structure of
606 California sea lions in the Gulf of California, Mexico. – J. Appl. Ecol. 47: 47-56.
607 Wood, S.N. (2006) Generalized Additive Models: An Introduction with R. Chapman
608 and Hall/CRC Press, Boca Raton, FL.

609 Figure Legends

610 Figure 1. Natural log of MASE statistics for 13 models, for prediction at t=1 to 4.
611 'Reg' = ordinary least-squares regression, 'MA' = moving averaged errors
612 ARIMA(0,0,1), 'RW' = random walk without drift , 'ARMA' = ARIMA(1,0,1) with a
613 constant, 'Exp' = exponentially smoothed ARIMA(0,1,1) , 'ARcor' = AR model with
614 temporally correlated errors (ARIMA(1,1,0)), 'ArSS' = state-space RW with drift
615 model, 'GAM' = generalized additive model, 'Loc' = weighted local regression , 'NN' =
616 neural network model, 'SMAP' = distance weighted non-parametric prediction, 'Smp' =
617 Simplex, 'RF' = random forest. Horizontal dashed lines correspond to the MASE from
618 the RW model without drift (RW-MASE). Number of time series for each dataset: n=214
619 (marine fish), n=289 (salmon), n=1322 (birds), n=46 (mammals). These models shown
620 were selected to summarize the overall behavior for model classes. The results for all
621 individual models are in Table S2.
622
623 Figure 2. Natural log of mean absolute square error (MASE) statistics for 13 models,
624 applied to different time series of salmon over prediction intervals 1 to 4. See Fig. 1
625 for the model descriptions for the model acronyms on the *x*-axis. Horizontal dashed
626 lines correspond to the MASE from the RW model. Number of time series for each
627 species: n=28 (pink, *O. gorbuscha*), n=40 (chum, *O. keta*), n=5 (coho, *O. kisutch*), n=61
628 (sockeye, *O. nerka*) and n=183 (Chinook, *O. tshawytscha*).
629
630 Figure 3. Biological effects of covariates (Table 2) that were correlated with changes
631 in the absolute scaled error (ASE) statistic from the GAM model, averaged over

632 forecasts of 1 to 3 time steps. The expected improvement in ASE is calculated as the
633 ASE statistic divided by the ASE statistic at the mean of each covariate (e.g. mean
634 trophic level of 2.5 for birds), $100 \times ASE_i / ASE_{\bar{x}}$. The solid line represents the
635 expected value, and the shaded region represents the 95% confidence intervals. The
636 darkness of the gray scale is proportional to the normal density.
637
638 Figure 4. Distribution of autocorrelation values for each of the datasets included in
639 our meta-analysis. These values represent the ACF at lag 1 of differenced values.

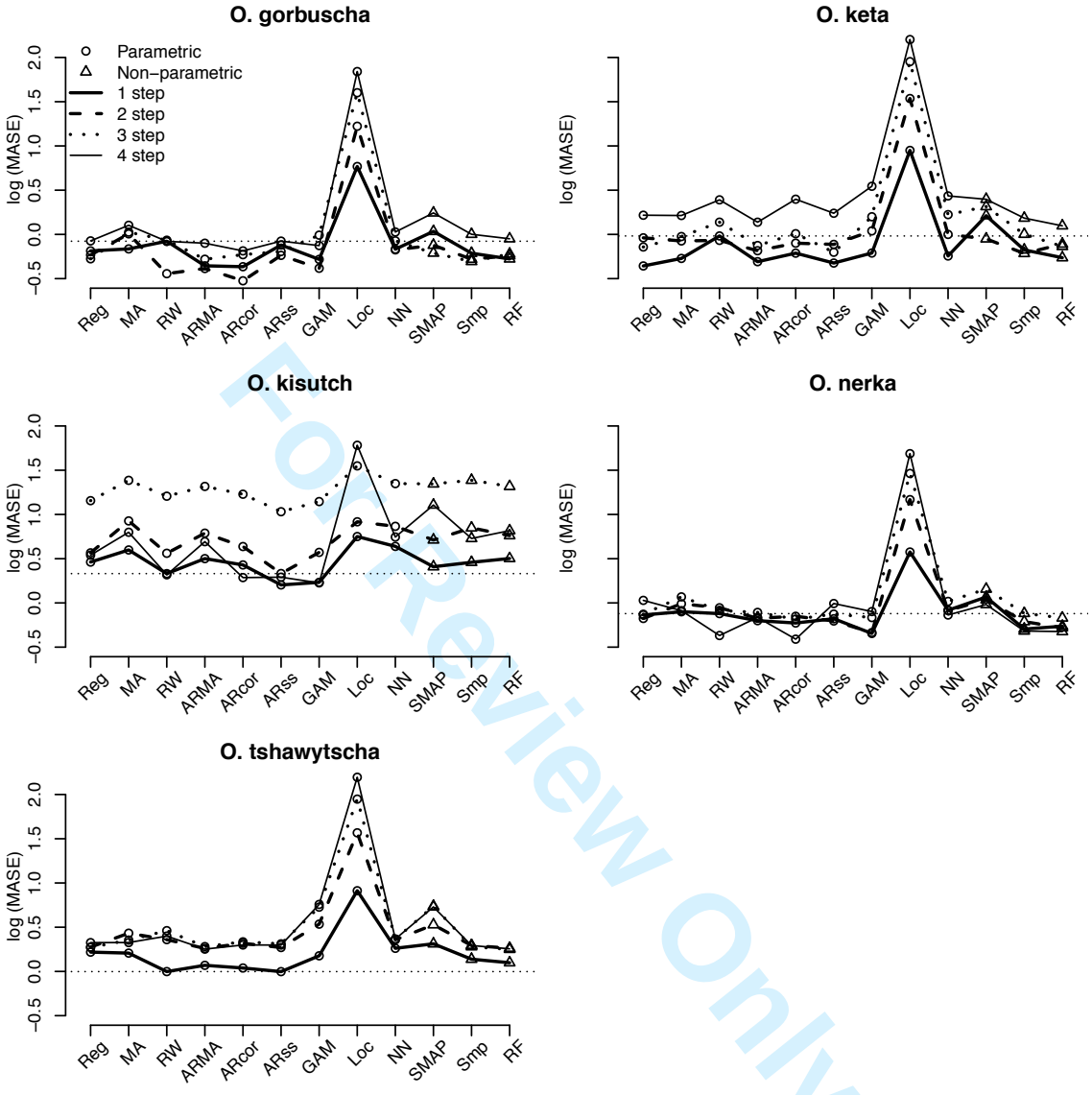
640 Figure 1.



641

642

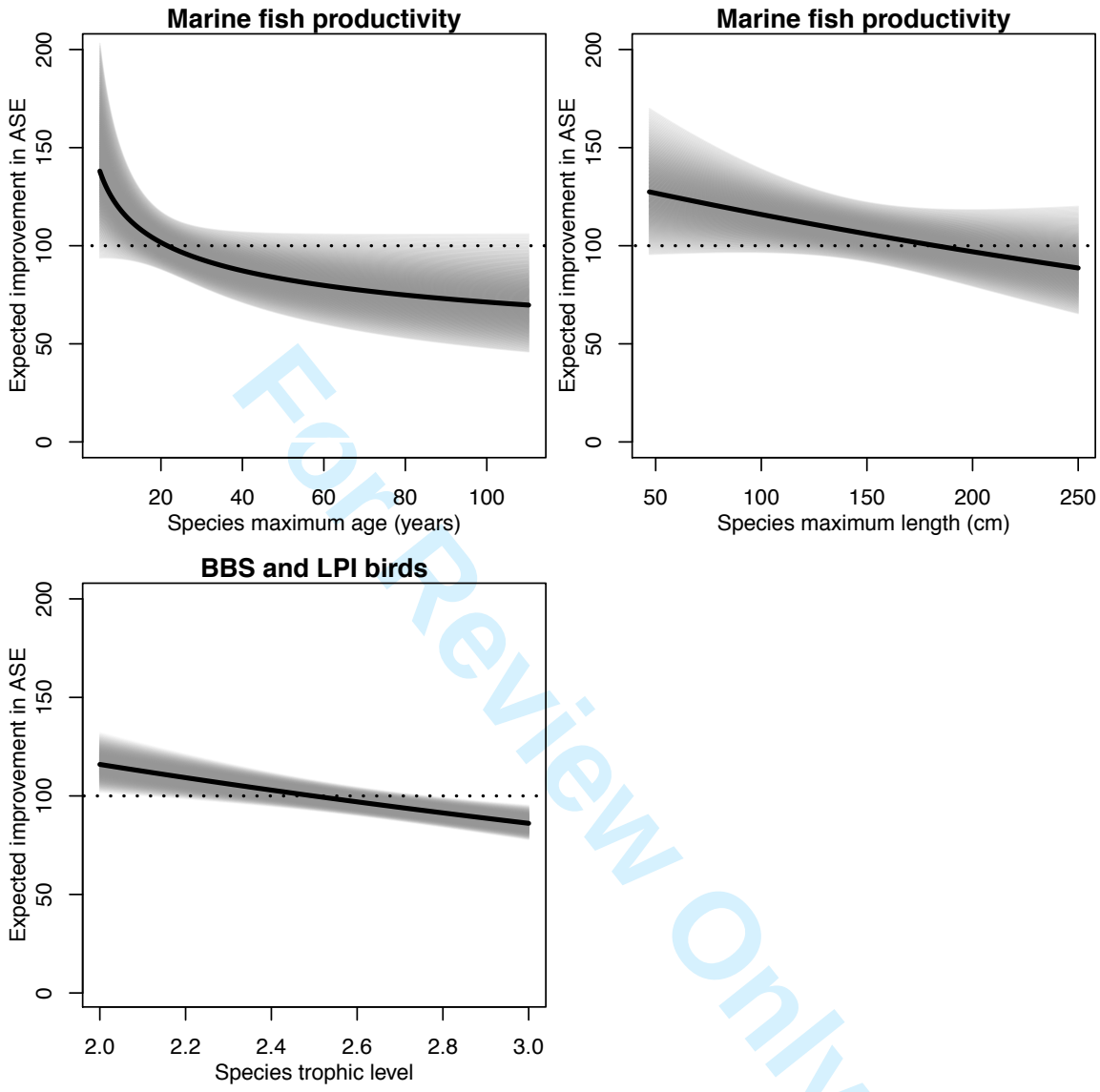
643 Figure 2.



644

645

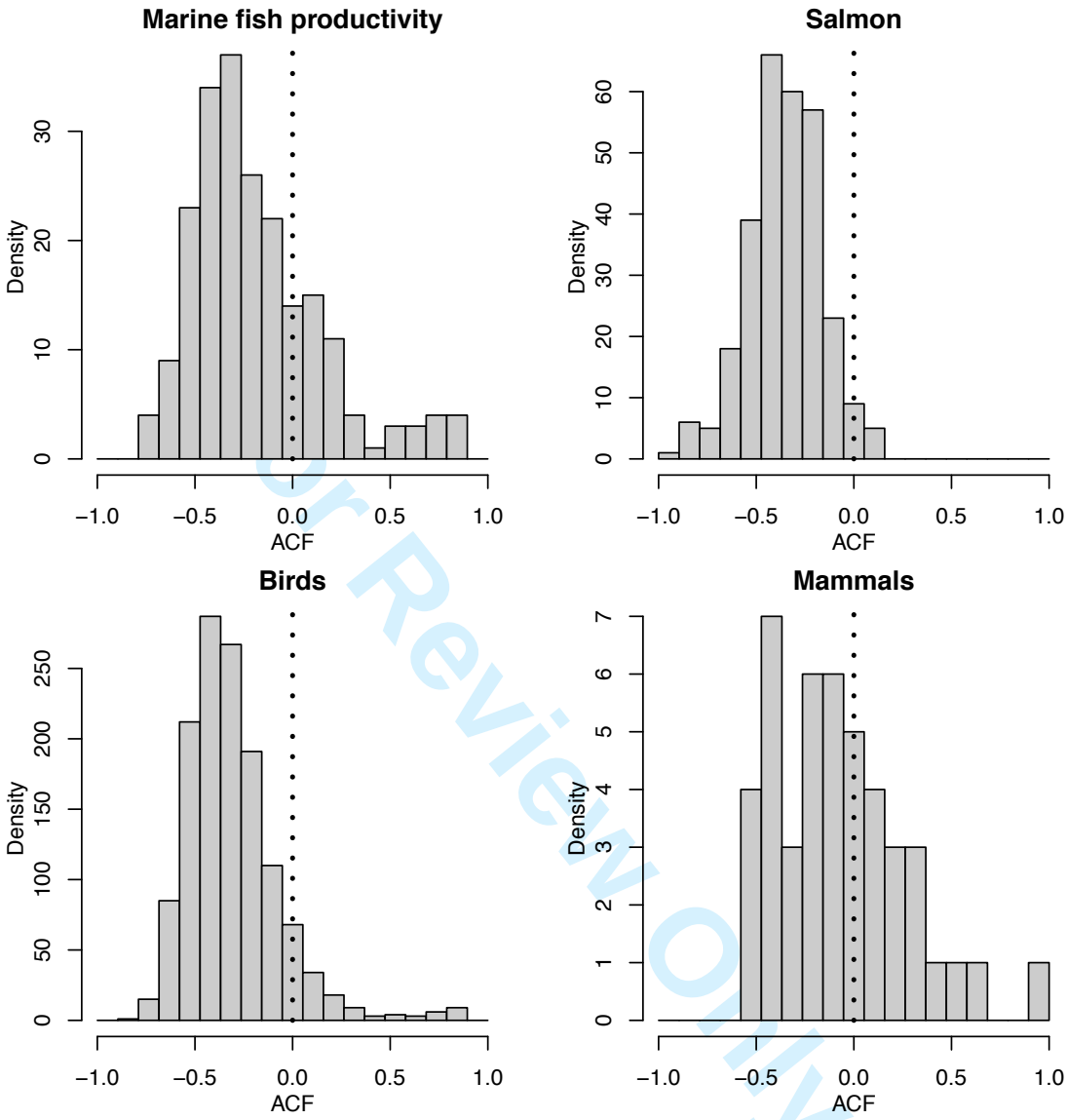
646 Figure 3.



647

648

649 Figure 4.



650

651

652 Table 1. Summary of time series datasets included in the meta-analysis

Dataset	Time series	Organism	Source
US BBS bird	414	Birds	Sauer *et al.* 2011
UK RSPB bird	61	Birds	Risely *et al.* 2012
LPI	1162	Birds, fish, mammals	Loh *et al.* 2005; Collen *et al.* 2009
RAM Recruits/spawner	214	Fish	Ricard *et al.* 2011
WA, OR salmon	44	Fish	Ford *et al.* 2010
CA salmon	155	Fish	Holmes & Fagan 2002
BC salmon	90	Fish	Dorner *et al.* 2008

653

Table 2. Regression parameters that have negative effects are associated with reduced MASE (improved forecasts over random walks). Regression coefficients are shown, with standard errors in parentheses. The quantity $\sigma^2_{obs}/\sigma^2_{pro}$ represents the ratio of observation to process variance, σ^2 represents the total variance of the time series deviations $(Y_{i+1} - Y_i)$ within the training data, and $\sqrt{|\rho|}$ represents the square root of the lag-1 autocorrelation in the raw training data.

Fish		Birds	
ln (age)	-0.187 (0.111)	Trophic level	-0.092 (0.050)
ln (length)	-0.282 (0.152)	Ln (σ^2)	0.065 (0.187)
ln ($\sigma^2_{obs}/\sigma^2_{pro}$)	-0.012 (0.003)	$\sqrt{\|\rho\|}$	-0.248 (0.117)

Supporting Information

Summary of models included

1. Random walk with drift

$Y_t = Y_{t-1} + u + e_t;\ e_t \sim Normal(0, \sigma)$

The drift term is *u*. This is a process error only model, with errors that are temporally independent.

2. Random walk with autocorrelated errors

$Y_t = Y_{t-1} + u + e_t;\ e_t \sim Normal(\rho \cdot e_{t-1}, \sqrt{1-\rho^2}\sigma)$

This is a process error only model, with errors that are temporally correlated ($-1 < \rho < 1$).

3. State space random walk model

Process equation: $X_t = X_{t-1} + u + e_t;\ e_t \sim Normal(0, \sigma)$

Observation (or 'data model') equation: $Y_t = X_t + \delta_t;\ \delta_t \sim Normal(0, \gamma)$

While the process model is a random walk, the total variance is broken up into a process component (representing natural stochasticity) and observation error component (resulting from imperfect observations and sampling error) (Lindley 2003).

4. Generalized additive models (GAMs)

Our implementation of GAMs only used time as a covariate, so the model was not autoregressive. The basic form is

$g(E[Y]) = B_0 + f(time)$

where the function *g*() is a link function (we used log), B_0 is an intercept, and the function *f*() is a smoothing function, or set of polynomial regression splines. The degree of smoothness was selected by cross validation (Wood 2006).

5. Neural network model

The neural network time series model is autoregressive, but non-linear,

$$Y_{t+d} = B_0 + \sum_{j=1}^{d} B_j g\left(\gamma_{0,j} + \sum_{i=1}^{m} \gamma_{1,j} \cdot Y_{t-(i-1)d}\right)$$

where the structure of the network is controlled by the embedding dimension (m) and time delay (d). The activation function $g()$ was assumed linear, and all other parameters represent weights or coefficients. Because of relatively short time series, we constrained m = 1:3, and d = 1:2.

6. ARIMA models

AR models treat x_t as autoregressive. The p term is the degree of lag included in the model:

AR: $Y_t = b_1 Y_{t-1} + b_2 Y_{t-2} + \ldots + b_q Y_{t-p} + e_t;\ e_t \sim Normal(0, \sigma)$

MA models have treat the errors, e_t, as autoregressive. The q term is the degree of lag included in the autoregressive model for the errors. A MA model with no AR component would be:

MA: $Y_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q};\ e_t \sim Normal(0, \sigma)$

An ARMA model is a time series model with both the AR and MA components. ARMA models may also include a constant. For example, AR(1) with constant would be

AR(1)+constant: $Y_t = b_1 x_{t-1} + \mu + e_t;\ e_t \sim Normal(0, \sigma)$

If b_1 is set to 1, this is a random walk with drift.

An ARIMA model includes both the AR and MA components but also specifies whether the raw data, Y_t, or lag-d differences are being modeled. An ARIMA model is denoted ARIMA(p, d, q). Thus a ARIMA(0,2,1) model would mean:

ARIMA(0,2,1): $Y_t - Y_{t-2} = e_t + \theta_1 e_{t-1};\ e_t \sim Normal(0, \sigma)$

It should be noted that most ARIMA models---the random walk with drift model being a major exception---are stationary, meaning they do not have a long-term temporal trend. When the time series has a trend, ARIMA models are used to model the residuals of a regression of that time series. We used the `Arima()` function in the forecast package in R which takes care of estimating the linear trend and fitting the residuals with the specified stationary ARIMA model. This can also be done using the base `arima()` function in R by passing in `xreg=1:n` as a covariate.

7. Exponentially smoothed time series

The most basic exponentially smoothed (or weighted) moving average time series models are ARIMA(p = 0, d = 1, q = 1),

$z_t = \sum_{j=1}^{\infty}(1-\lambda)\lambda^{j-1} z_{t-j} + e_t;\ e_t \sim Normal(0, \sigma);\ |\lambda| < 1$ (Shumway & Stoffer 2006)

Where z_t is the detrended data, Y_t-(a-bt), and a+bt is the linear trend (estimated simultaneously with the ARIMA model for the residuals).

8. Local regression

Local regression represents a linear model that is fit piecewise, in a moving window procedure, through a time series, and the prediction at a given time point is a function of data in the past and future,

$$\hat{Y}_t = f(Y) + e_t;\ \ e_t \sim Normal(0, \sigma)$$

The function *f*() typically takes two arguments: a nearest neighbor or bandwidth argument, specifying how much of the dataset to use (0-100%), and a parameter or function controlling the exponential decay between points. For each dataset in our analysis, we used cross validation to select the nearest neighbors and polynomial (1:3). The parametric version of this model was implemented using locfit(), and a non-parametric version of the model was implemented with a kernel regression estimator using the npreg() function.

9. Gaussian process regression

The objective of Gaussian process regression is to make prediction while conditioning on a covariance matrix, $\mathbf{\Sigma}$, and previously observed residuals.

$$\hat{Y}_t = f(Y) + e_t;\ \ e_t \sim Multivariate\ normal(0, \mathbf{\Sigma})$$

All data points are assumed to have arisen from an unknown covariance function, and unlike other methods (e.g. local or non-parametric bandwidth regression), the correlation between points is not modeled as a function of the distance between them in time, but in terms of their relative values (e.g. biomass or abundance at time *t* and *t*+1).

10. Random forest regression

Random forest uses an ensemble prediction from n_{trees} different regression trees (we have used n_{trees} = 500). Each tree uses a bootstrap of the data, and a randomly chosen subset of the predictor variables. This is done to minimize the correlation among predictions from different trees, which will tend to decrease predictive error for ensemble forecasting methods. For predictor variables we have used a basis-expansion using the lag-operator, and lags 1-10.

$$\hat{Y}_t = \frac{1}{n_{trees}} \sum_{i=1}^{n_{trees}} \hat{Y}_{t,i}$$

where $\hat{Y}_{t,i}$ is the prediction from the *i*-th tree. Each tree starts with the following prediction:

$$\hat{Y}_t = \frac{1}{n} \sum_{j=1}^{n} Y_j$$

The tree then searches among available variables and finds the variable and split that maximizes the reduction in root-mean-squared error. This process is repeated until a particular node has 5 or fewer observations.

11. Simplex

The goal of simplex is to predict the dynamics of a variable without using a parametric equation, and hence potentially avoiding problems associated with parametric models that occur when dynamics are highly state-dependent. Simplex does this by identifying nearest neighbors using a Euclidean distance metric defined in a *d*-dimension space generated using the lag-operator.

$$\hat{Y}_t = \frac{1}{d+1} \sum_{i=d+1}^{t-f} I(D_i) \cdot Y_i$$

where d is the embedding dimension, f is the prediction interval, D_i is a Euclidean distance in d-dimensional lag-space:

$$D_i = \sqrt{\sum_{j=1}^{d} \left(Y_{i-j} - Y_{t-j}\right)^2}$$

and $I(Y_{i-d},\ldots,Y_{i-1})$ is an indicator variable that identifies $d + 1$ nearest neighbors in the Euclidean distance D_i, i.e., equals one if distance D_i is one of the $d + 1$ lowest distances. The embedding dimension d is then selected using cross-validation.

12. S-MAP

S-MAP has a similar goal to Simplex, and typically uses the embedding dimension previously selected using Simplex. However, it has an additional parameter θ representing the degree of state-dependent dynamics in a time series. Instead of nearest neighbors, it calculates a weight γ_i for each point i using the distance defined for Simplex:

$$\gamma_i = \theta \cdot \frac{D_i}{\sum_{j=1}^{n} D_i}$$

This weight is then used to take a weighted average of the dynamics of all points.

$$\hat{Y}_t = \langle 1, Y_{t-f}, \ldots, Y_{t-f-d} \rangle \times \boldsymbol{C}$$

where × is the matrix multiplicative operator and $\boldsymbol{C}$ is the solution to a weighted linear model:

$$\boldsymbol{C} = \boldsymbol{A}^{-1} \times \boldsymbol{B}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are formed from the lagged variables, and the inverse of A is accomplished using the singular-value decomposition:

$$\boldsymbol{B} = \boldsymbol{\gamma} \cdot \boldsymbol{x}_{-t}$$

where · is the pairwise multiplication operator and $\boldsymbol{x}_{-t}$ is the vector of the time series excluding observation x_t, and

$$\boldsymbol{A} = \langle \boldsymbol{\gamma} \cdot 1, \boldsymbol{\gamma} \cdot l_f(\boldsymbol{Y}_{-t}), \ldots, \boldsymbol{\gamma} \cdot l_{f-d+1}(\boldsymbol{Y}_{-t}) \rangle$$

and $l_f(\boldsymbol{Y}_{-t})$ is the lag operator of order f for the vector $\boldsymbol{Y}_{-t}$.

Table S1. Model summary and the code / functions used to fit them in existing packages in the R programming environment.

Model	R package (R function in package)	Parametric
Random walk	forecast (rwf)	Y
State-space random walk	stats (StructTS), MARSS (MARSS)	Y
GAMs	mgcv (gam)	Y
Neural network time series	tsDyn (nnetTs)	N
Exponentially smoothed time series	forecast (ets)	Y
Local regression	locfit (locfit)	Y
Kernel / bandwidth regression	np (npreg)	N
ARIMA	forecast (Arima), stats (arima)	Y
Gaussian process	kernlab (gausspr)	N
Random Forest	randomForest (randomForest)	N
SMAP, Simplex	Code by Jim Thorson; https://r-forge.r-project.org/R/?group_id=1316	N

Table S2. Table of 1-step ahead MASE statistics for 49 models in our analysis. R packages and functions used are listed in Table S1. Stationary ARIMA models (those not denoted RW), are fit to detrended data, but the forecast from those models includes the trend.

Model	Marine fish Productivity	Salmon	Birds	Mammals
GAM (gam)	1.768	1.040	0.969	1.087
neural network (1,1)	1.850	1.152	1.420	2.191
neural network (1,2)	1.736	1.222	1.197	1.560
neural network (2,1)	1.729	1.171	1.418	2.258
neural network (2,2)	2.109	1.273	1.217	1.451
neural network (3,1)	1.788	1.199	1.434	1.815
neural network (3,2)	2.093	1.413	1.297	1.720
RW no drift - ARIMA(0,1,0) without constant	1.431	0.982	0.976	1.062
RW with drift - ARIMA(0,1,0) with constant	1.449	0.994	0.994	1.159
Exp smooth with trend, ARIMA(0,1,1)	1.471	0.957	0.932	1.277
Exp smooth without trend, ARIMA(0,1,1)	1.473	0.966	0.940	1.277
Structural time series (freq=1)	1.429	0.905	0.904	1.136
Structural time series (freq=2)	1.474	0.962	0.940	1.151
Local regression	2.490	2.333	1.940	2.356
Kernel/bandwidth regression	1.545	1.018	0.961	1.146
ARIMA(1,0,1)	1.414	0.965	0.986	1.175
Gompertz; ARIMA(1,0,0)	1.381	0.976	1.037	1.091
ARIMA(2,0,1)	1.430	0.997	1.000	1.212
ARIMA(1,0,2)	1.478	1.027	1.009	1.136
ARIMA(2,0,2)	1.481	1.021	1.005	1.212
MA model; ARIMA(0,0,1)	1.731	1.118	2.112	1.711
ARIMA(0,0,2)	1.695	1.068	1.715	1.477
ARIMA(2,0,0)	1.386	0.993	1.005	1.175
ARIMA(1,1,1)	1.414	0.913	0.915	1.164
ARIMA(1,1,0)	1.399	0.942	0.933	1.103
ARIMA(2,1,1)	1.407	0.936	0.920	1.214
ARIMA(1,1,2)	1.426	0.935	0.923	1.215
ARIMA(2,1,2)	1.445	0.981	0.951	1.217
ARIMA(0,1,1)	1.422	0.893	0.911	1.174
ARIMA(0,1,2)	1.455	0.934	0.934	1.205
ARIMA(2,1,0)	1.402	0.940	0.923	1.208
ARIMA(1,2,1)	1.421	0.958	0.907	1.189
ARIMA(1,2,0)	1.731	1.279	1.208	1.290
ARIMA(2,2,1)	1.422	0.965	0.910	1.173
ARIMA(1,2,2)	1.445	0.950	0.901	1.295
ARIMA(2,2,2)	1.452	0.963	0.936	1.183
ARIMA(0,2,1)	1.435	0.994	0.967	1.191
ARIMA(0,2,2)	1.476	0.901	0.897	1.240
ARIMA(2,2,0)	1.626	1.183	1.107	1.269
Gaussian process (freq=1)	1.691	1.042	1.730	1.597

Gaussian process (freq=2)	1.716	1.014	1.706	1.570
Gaussian process (freq=3)	1.749	1.014	1.731	1.396
Gaussian process (freq=4)	1.743	1.029	1.706	1.586
State-space RW with drift	1.482	0.928	0.966	1.295
State-space RW no drift	1.464	0.909	0.915	1.155
Simplex	1.578	0.990	1.337	1.321
S-MAP	1.658	1.291	1.483	2.156
Random Forest regression	1.562	0.988	1.124	1.197
linear regression	1.886	1.094	1.549	1.925