The intrinsic predictability of ecological time series and its potential to guide forecasting

Reproducible report of statistical analysis of empirical time series

Frank Pennekamp

17 December 2018

Contents

Motivation

Our conceptual framework as well as our simulation results suggest a relationship between time series complexity as measured by permutation entropy (PE) and forecasting error. However, the simulations have also shown that problems commonly encountered by ecologists such as short and noisy time series can considerably distort the relationship. Therefore our first goal was to test for such a relationship across a range of ecological time series, and our second goal was to test whether time series covariates such measurement error (i.e. lab versus field time series), the non-linearity of the system or the time series length can explain deviations from a general relationship. To do so, we relied on a freely available compilation of time series from the Global Population Dynamic Database and a second database of laboratory based time series for which time series complexity was estimated by PE. We also forecasted the time series using the Empirical Dynamic Modelling (EDM) approach developed by George Sugihara and co-workers and made available via the rEDM package. This notebook documents this analysis.

Load and prepare data for analysis

GPDD

The GPDD data was accessed via the rGDPP package. The database contains almost 5000 time series that vary widely in length and quality and is one of the go-to databases for ecologists to ask broad population dynamic questions (Inchausti & Halley 2001). The rGPDD package contains tables with the time series information and additional information about the data sources, measurement units, and some fitted population dynamic parameters. For our purpose, we mainly use the "data" table, but added information about the source to control for potential non-independence in subsequent analyses. The plot shows a random sample of 12 time series from the GPDD.
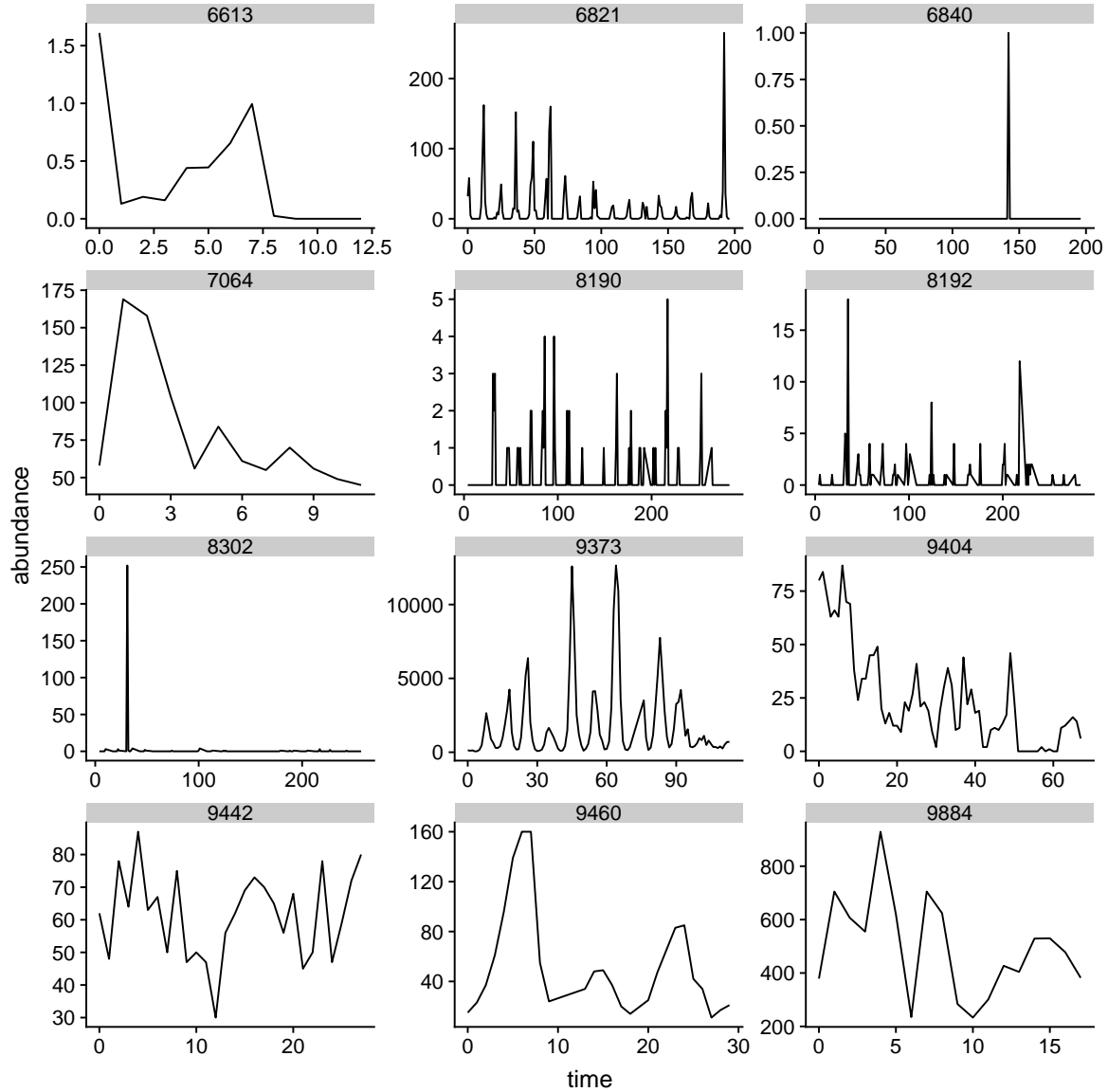

Laboratory data

For the database of laboratory time series we compiled data from Hiltunen et al. 2014, Fussmann et al. 2000 and Becks et al. 2005. Whereas the first publication is itself a compilation of laboratory time series,

the second and third study used two systems (microbial predator-prey systems) across a range of chemostat dilution regimes. The two papers report how the different dilution regimes push the systems through a range of dynamical regimes spanning steady-state dynamics till chaotic dynamics. The following three plots show the dynamics across the systems. For the analysis we only used information from the univariate time series, although EDM would also allow to incorporate information the other species in the system to improve forecasts.
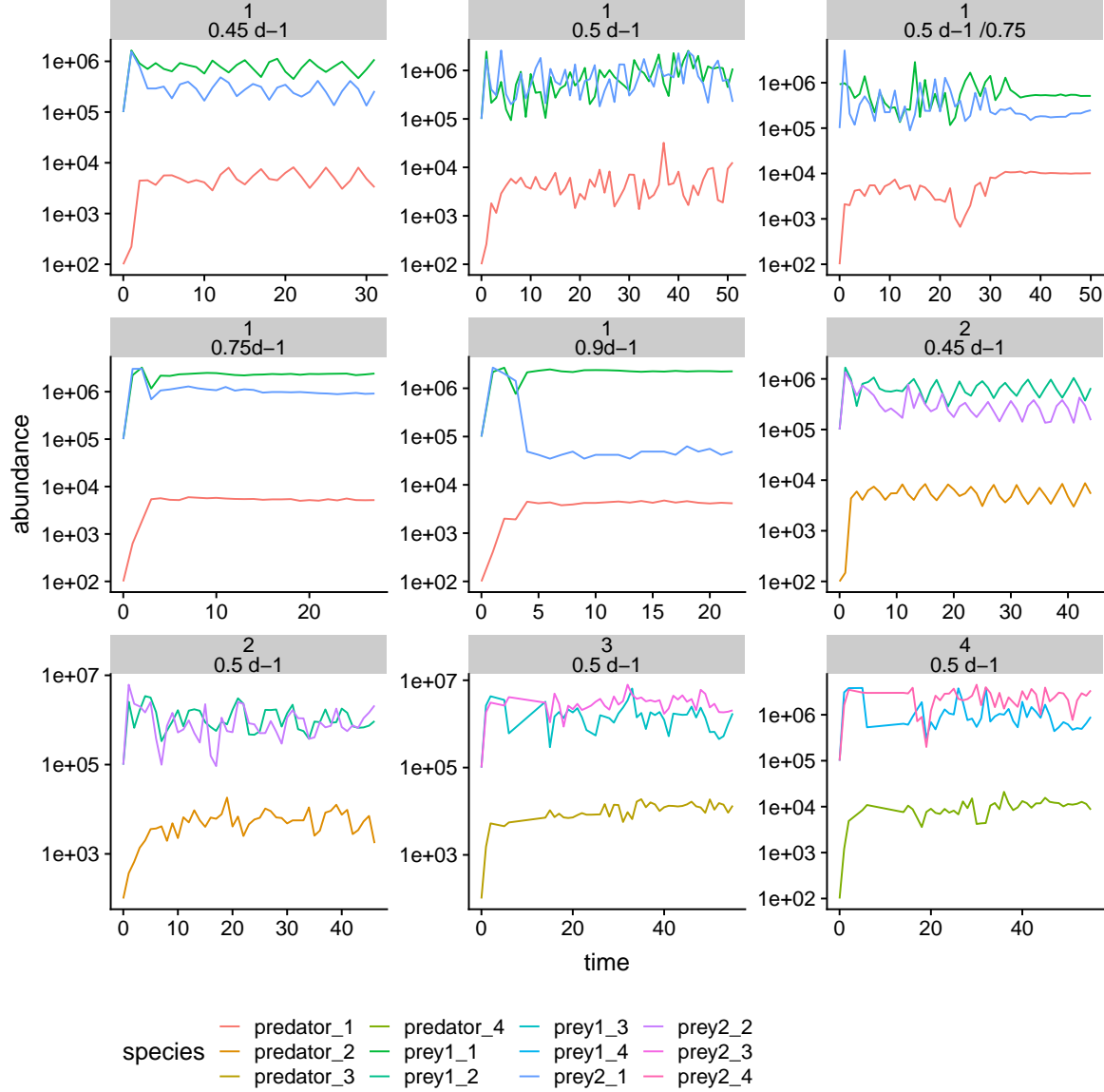
Hiltunen et al. 2014 data

species — predator — prey
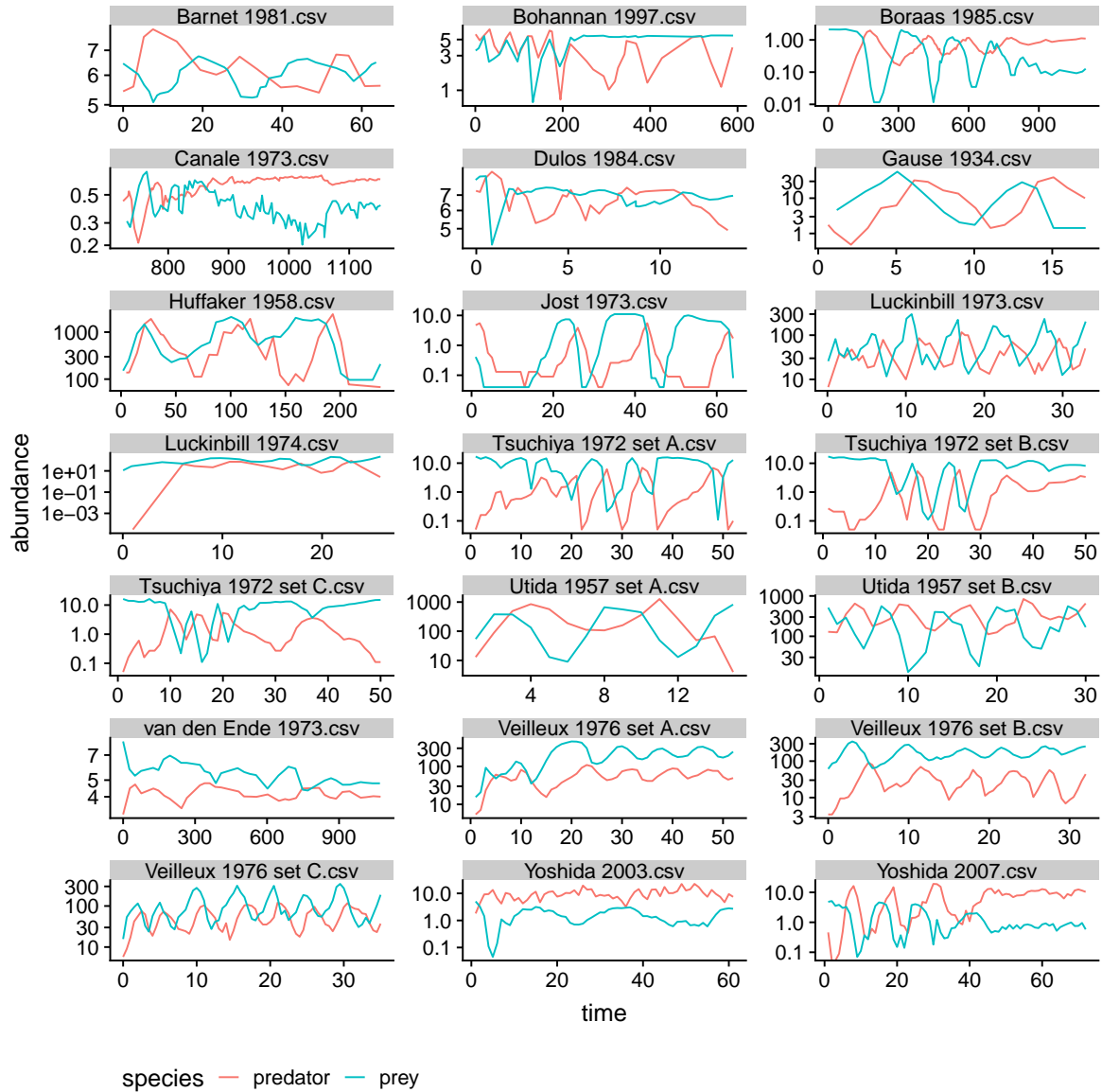
Fussmann et al. 2000 data
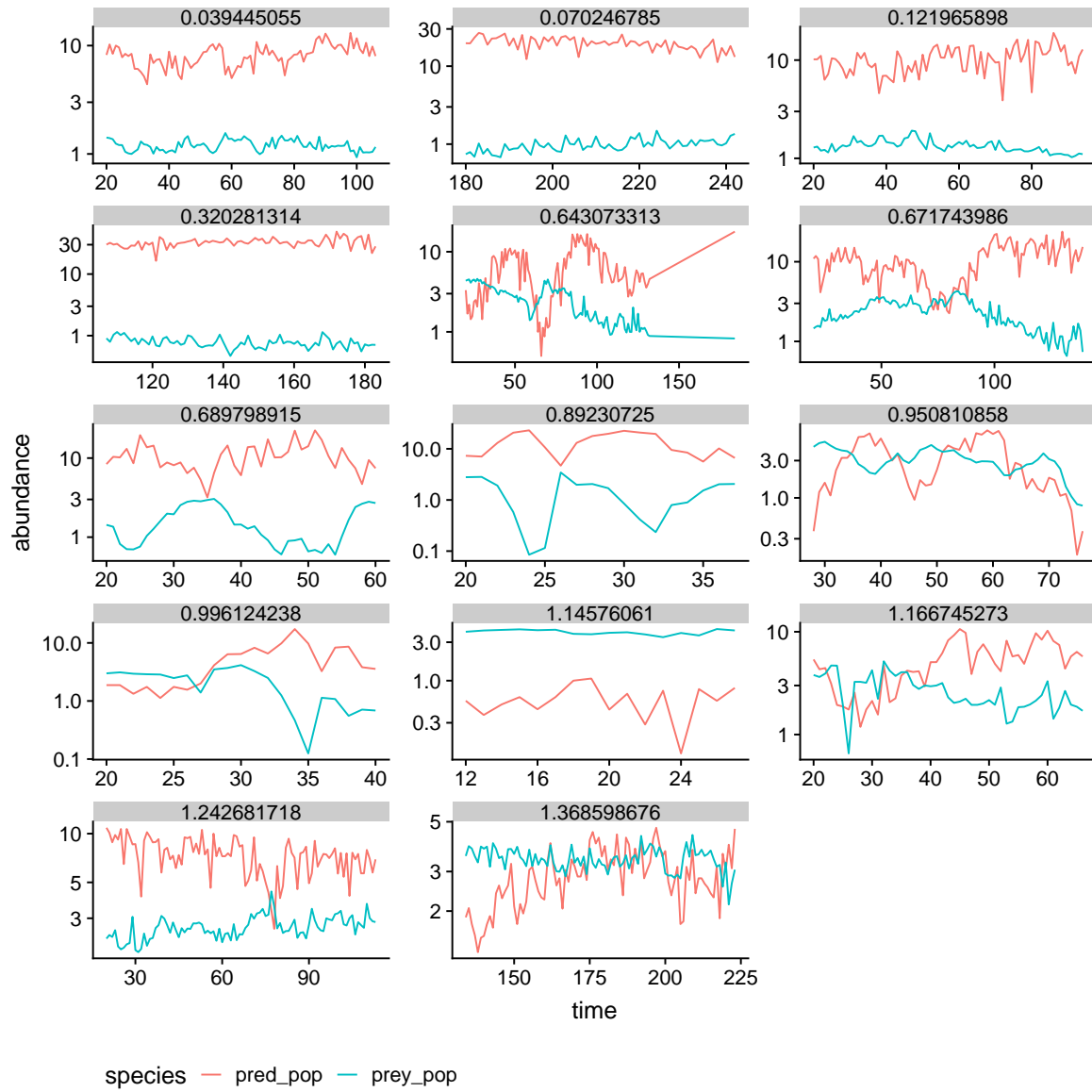
Becks et al. 2005 data

Data transformation

No data transformation applied. Centering of time series did not qualitative change the results.

Filtering datasets for analysis

We extracted some descriptive information about each of the time series. We calculate a number of time series properties for filtering: number of observations (N), number of zeros (N_zeros), proportion of zeros in the time series (zero_prop) and proportion of ties in the time series (ties_prop). Zero and tie proportions were calculated as they pose problems to the rank-based PE. However, different methods to deal with ties are available.

We decided to only use time series with a minimum number of 30 observations and no gaps. This number is the minimum, as for very short time series PE may not be meaningful and also forecasting via EDM is

limited. We also specified that the proportion of ties in the time series should be less than 15%.

Adding NAs for missing data

Time series were padded with NAs whenever there were missing measurements.

Calculate permutation entropy across time series

We calculated the weighted permutation entropy for all the time series using a word length of 3 and tau of 1. Other values for word length and tau did not change the results. Moreover, using other tie breaking methods (e.g. average or random) did not change the results.

Distribution of PE values over lab and field data

The following graph shows the distribution of permutation entropy values for field (GPDD) and lab studies.
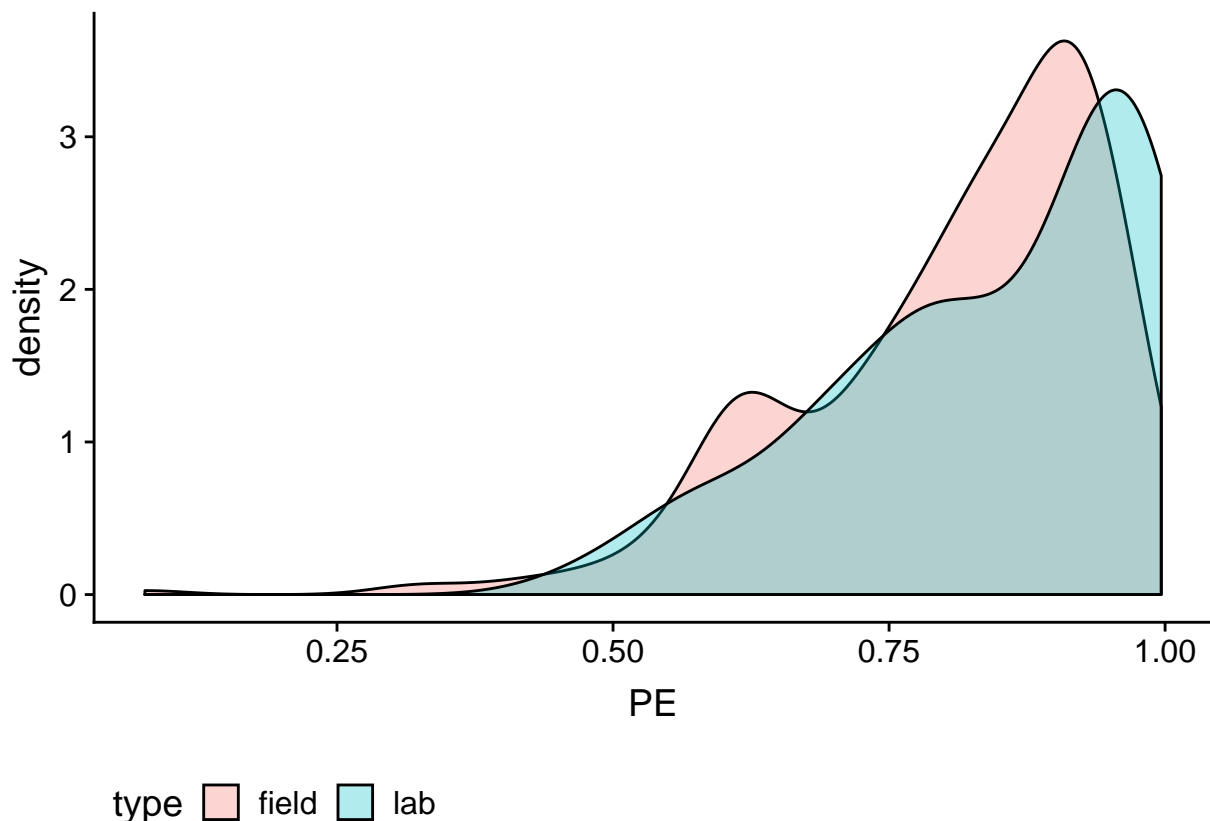
Forecasting via simplex and s-map projection (using a split dataset)

We forecasted each individual time series using the Empirical Dynamic Modelling (EDM) We divided the time series into a training and a test dataset. 2/3 of the time series was used for training, and the remaining 1/3 for testing.

The EDM approach consists of two steps to determine the embedding dimension E and the tuning parameter theta that determines the degree of non-linearity of the forecast. The embedding dimension determines the

number of lags used for reconstructing the state-space and the tuning parameter theta determines how non-linear the system is (i.e. which of the neighboring points in the state-space are considered for forecasting).

We first fitted the simplex projection to identify the best embedding dimension E on the training data. Usually dimensions between one and ten are tested and the dimension with the highest forecast skill using a leave-one-out-cross validation is chosen. In the second step, we fitted the S-maps with the determined dimension E to the training data and tested the tuning parameter theta. Theta varied in 18 steps between 0 and 8 (with log scaled intervals). Finally, we tested the final EDM model on test part of each time series and recorded the forecasting skill.

```
## pdf
##   2
```

Extract and plot information about the fitted EDM models (embedding dimension and theta)

Both embedding dimension and theta are additional features of the time series and may have explanatory value. We therefore added these metrics to our data.frame to control for the parameters in subsequent analyses. The two following plots show the distributions of D and theta across the two databases.
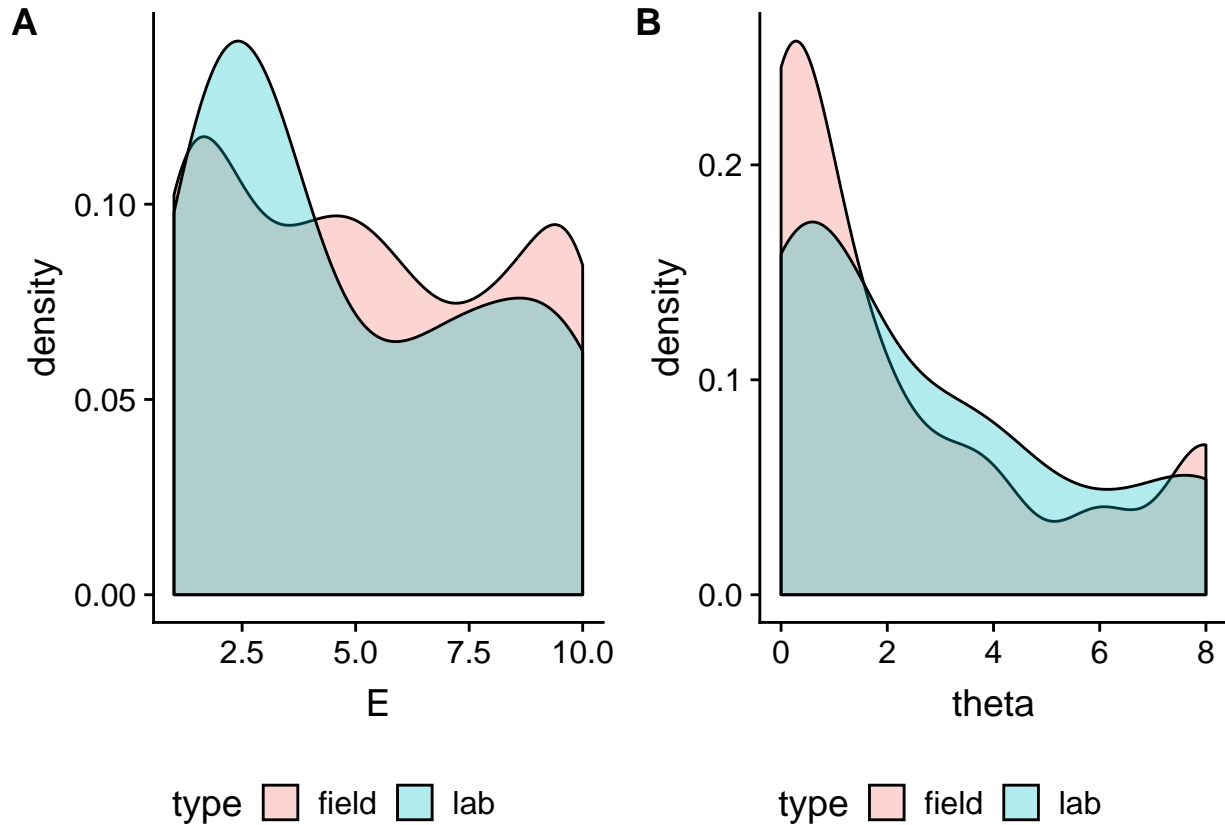
Calculate forecast error metrics

To evaluate the forecasting error, we used the normalised RMSE (nRMSE) as an appropriate measure to compare across time series.

Descriptive information about time series included

Skim summary statistics

n obs: 461

n variables: 32

Variable type: integer

variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100	hist
E	0	461	461	5.18	3.1	1	2	5	8	10	
N	0	461	461	62.61	38.04	30	39	50	63	197	

Variable type: numeric

variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100	hist
nrmse	0	461	461	0.21	0.14	9.3e-06	0.12	0.19	0.27	1.37	
PE	0	461	461	0.81	0.14	0.076	0.73	0.84	0.92	1	
theta	0	461	461	2.39	2.78	0	0	1	4	8	
ties_prop	0	461	461	0.018	0.031	0	0	0	0.026	0.15	
zero_prop	0	461	461	0.014	0.044	0	0	0	0	0.3	

Is there a relationship between PE and FE? Which of the time series covariates explain deviations from the relationship?
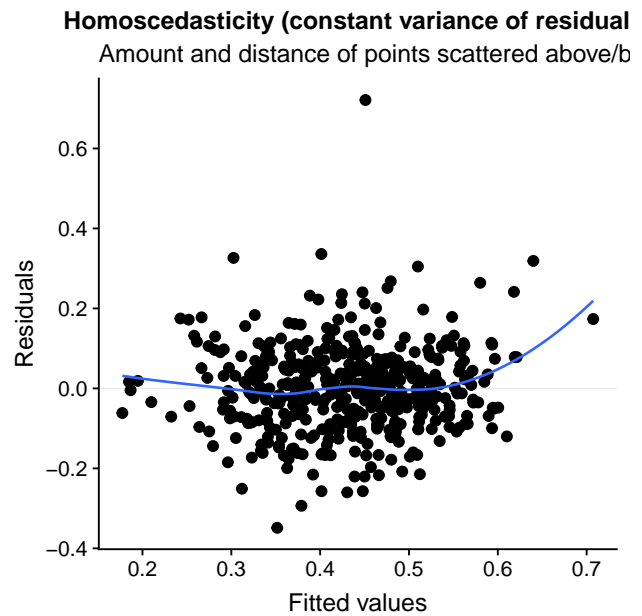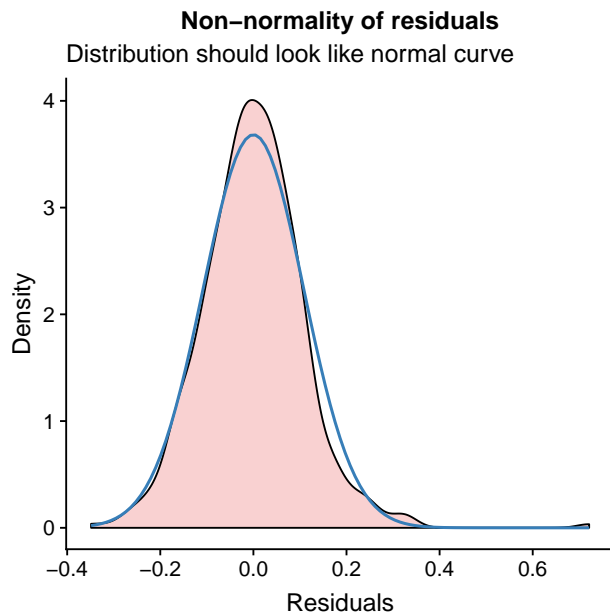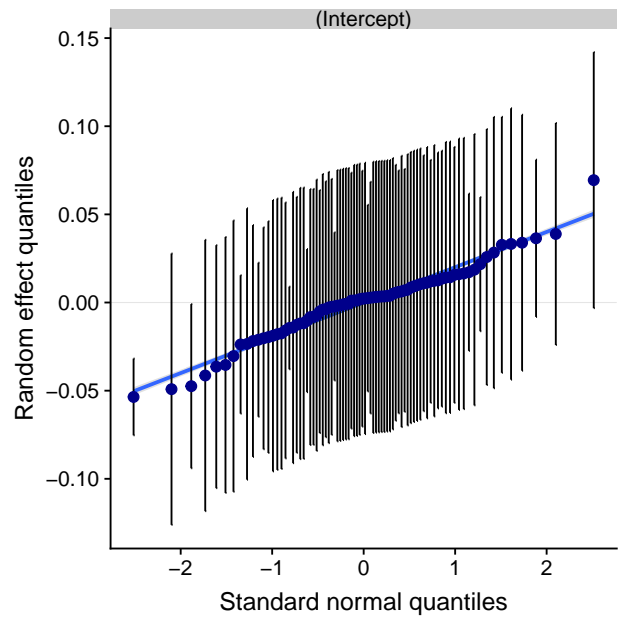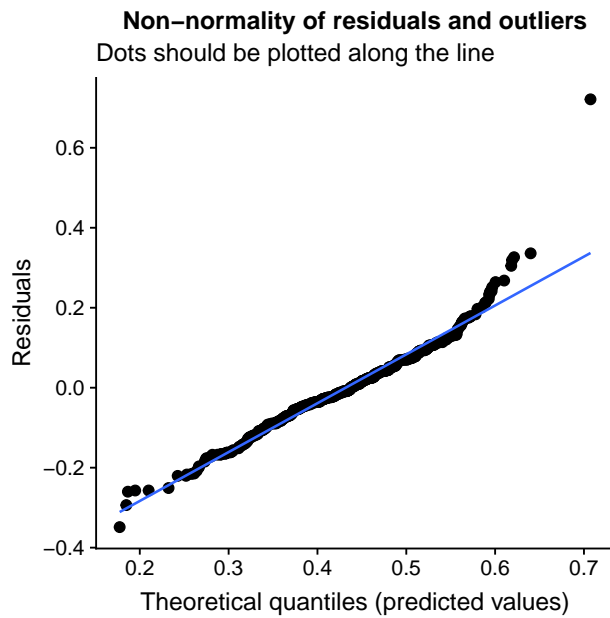
For analysis and plotting we centered the PE values to test for differences in intercept in the mean range of PE values rather than at the theoretical limit of 0 (perfect intrinsic predictability).

We find a relationship between PE and nRSME. Forecast error increases with time series complexity (i.e. increasing PE). The analysis using a mixed model with PE and type as fixed effect and time series source as random effect shows that the PE slope is significantly different from zero. Neither type (i.e. field versus lab time series) nor the interaction with PE is significant. Additionally, we were interested in whether time series properties such as length or non-linearity can explain deviations from the general relationship. We found that in addition to PE, time series length (N), the proportion of zeros (zero_prop), time series nonlinearity (theta) and dimensionality (E) had significant effects on nRMSE.

|	Estimate	CI (lower)	CI (upper)	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	0.0894081	0.0090585	0.1655839	0.0401685	364.9241	2.2258240	0.027	*
PE	0.4794040	0.3942410	0.5633243	0.0434695	451.5629	11.0285214	<0.001	***
type: lab	-0.0682111	-0.2858519	0.1557931	0.1144166	284.8046	-0.5961645	0.552	
N	-0.0016938	-0.0021204	-0.0012548	0.0002223	50.4175	-7.6198619	<0.001	***
zero_prop	0.4125419	0.0889610	0.7525901	0.1705308	392.6970	2.4191634	0.016	*
ties_prop	-0.3487946	-0.7797260	0.0783061	0.2218958	382.0995	-1.5718845	0.117	
E	0.0087928	0.0052394	0.0124925	0.0018547	448.1884	4.7409571	<0.001	***
theta	0.0113275	0.0073399	0.0154709	0.0020816	449.6891	5.4417971	<0.001	***
PE:typelab	0.0921348	-0.1816177	0.3556121	0.1391723	207.4432	0.6620196	0.509	

Model diagnostics

The model diagnostics show that residuals are normally distributed and have homogenous variance. Equally, random effects are normally distributed around zero as expected.

Non−normality of residuals and outliers
Dots should be plotted along the line

(Intercept)

Non−normality of residuals
Distribution should look like normal curve

Homoscedasticity (constant variance of residual
Amount and distance of points scattered above/b

Calculate partial R squared to assess explanatory power of individual predictors

Effect	Rsq	lower.CL	upper.CL
Model	0.346	0.290	0.416
PE	0.208	0.148	0.273
N	0.181	0.123	0.244
theta	0.057	0.023	0.104
E	0.044	0.015	0.086
zero_prop	0.013	0.001	0.041
ties_prop	0.005	0.000	0.027
PE:typelab	0.001	0.000	0.015
typelab	0.001	0.000	0.015

Plot PE vs FE relationship and highlight its potential to diagnose specific time series
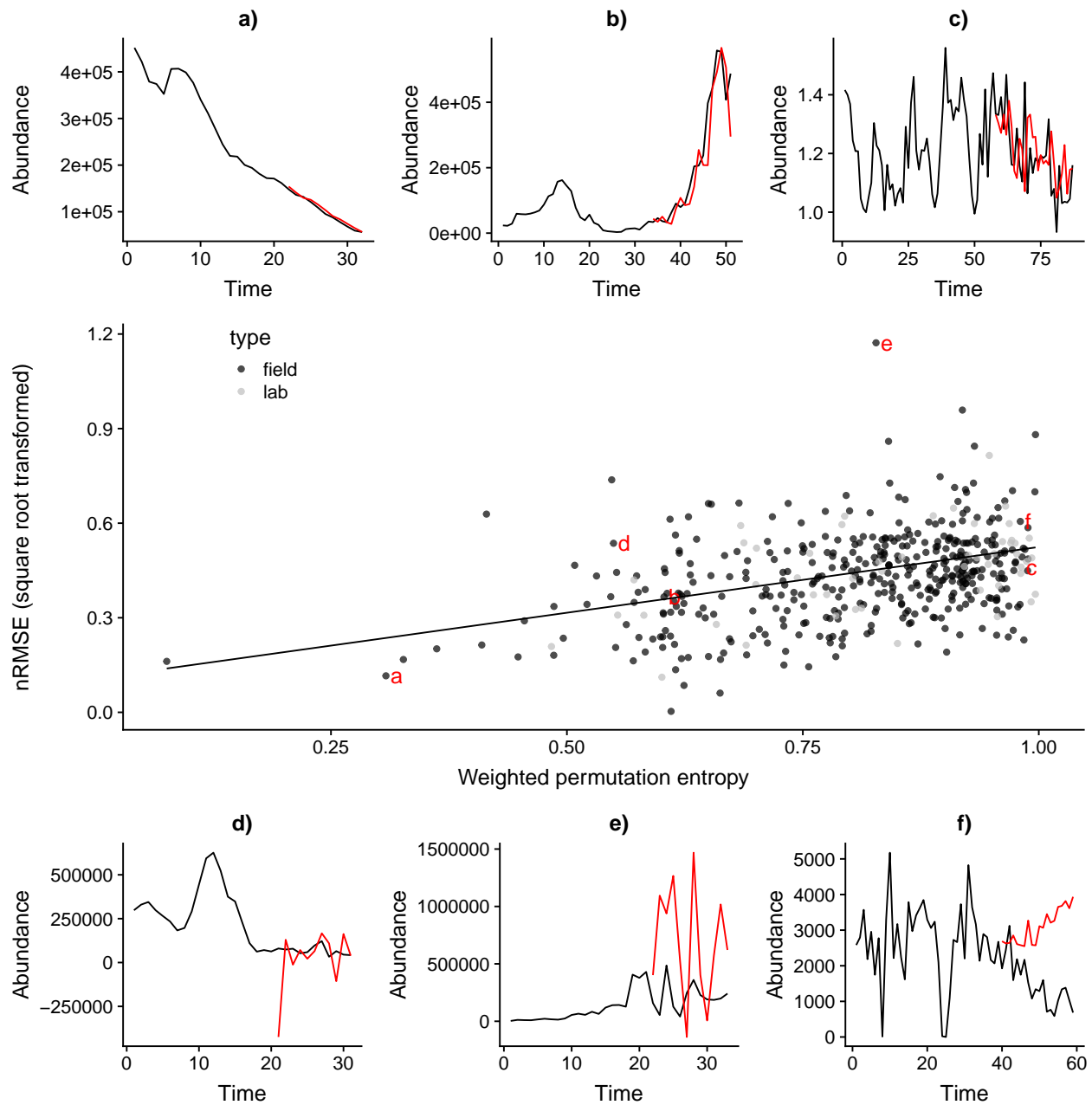

```
## pdf
##    2
```

References

Becks, L., Hilker, F.M., Malchow, H., Jürgens, K., Arndt, H., 2005. Experimental demonstration of chaos in a microbial food web. Nature 435, 1226–1229. https://doi.org/10.1038/nature03627

Fussmann, G.F., Ellner, S.P., Shertzer, K.W., Hairston, N.G., 2000. Crossing the hopf bifurcation in a live predator-prey system. Science 290, 1358–1360.

Hiltunen, T., Hairston, N.G., Hooker, G., Jones, L.E., Ellner, S.P., 2014. A newly discovered role of evolution in previously published consumer–resource dynamics. Ecology letters 17, 915–923.

Inchausti, P., Halley, J., 2001. Investigating Long-Term Ecological Variability Using the Global Population Dynamics Database. Science 293, 655–657. https://doi.org/10.1126/science.293.5530.655