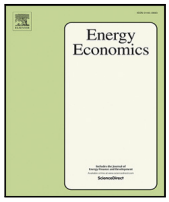
Energy Economics 124 (2023) 106833

Contents lists available at ScienceDirect

Energy Economics

journal homepage: www.elsevier.com/locate/eneeco



Assessing and comparing fixed-target forecasts of Arctic sea ice: Glide charts for feature-engineered linear regression and machine learning models

Francis X. Diebold [a,1], Maximilian Göbel [b,1], Philippe Goulet Coulombe [c,*,1]

[a] University of Pennsylvania, United States of America
[b] Bocconi University, Italy
[c] Université du Québec à Montréal, Canada

ARTICLE INFO

JEL classification:
Q54
C22
C52
C53

Keywords:
Seasonal climate forecasting
Forecast evaluation and comparison
Prediction

ABSTRACT

We use "glide charts" (plots of sequences of root mean squared forecast errors as the target date is approached) to evaluate and compare fixed-target forecasts of Arctic sea ice. We first use them to evaluate the simple feature-engineered linear regression (FELR) forecasts of Diebold and Göbel (2022), and to compare FELR forecasts to naive pure-trend benchmark forecasts. Then we introduce a much more sophisticated feature-engineered *machine learning* (FEML) model, and we use glide charts to evaluate FEML forecasts and compare them to a FELR benchmark. Our substantive results include the frequent appearance of predictability thresholds, which differ across months, meaning that accuracy initially fails to improve as the target date is approached but then increases progressively once a threshold lead time is crossed. Also, we find that FEML can improve appreciably over FELR when forecasting "turning point" months in the annual cycle at horizons of one to three months ahead.

1. Introduction

Arctic sea ice is melting very quickly as the planet warms (see Fig. 1, and e.g., Diebold and Rudebusch (2022) and the many references therein), which brings both major economic opportunities and major risks. Opportunities/benefits include new accessibility for extracting deposits of natural gas, petroleum, and other natural resources, as well as the emergence of trans-Arctic shipping lanes, which will enhance international trade by reducing both transportation costs and piracy-riddled chokepoints on other routes. Risks/costs include increased emissions and environmental damage due to discharges, spills, and soot deposits (Bekkers et al., 2016; Petrick et al., 2017). Finally and more broadly, melting sea ice will have important geopolitical consequences for Arctic sea-lane control (Ebinger and Zambetakis, 2009).

For all of the above reasons, the temporal path and pattern of Arctic sea ice diminution are of particular interest, and Arctic sea ice forecasting has received significant attention (Shalina et al., 2020). From a real-time online perspective, there are two key approaches. The first is fixed-horizon forecasting, where, for example, each month we forecast one month ahead, month after month, ongoing, as in Diebold and Rudebusch (2022). The second is fixed-target forecasting, where each month we forecast a fixed future target date, month after month, ending when we arrive at the target date, as in Diebold and Göbel (2022). In this paper we consider the fixed-target scenario, which has generated substantial interest in highlighting Arctic sea ice diminution both within years (as September 30 is approached, say) and across years (comparing the sequence of Septembers, say).[2]

Forecast accuracy naturally increases as information accumulates and the target date is approached. A key question is how to quantify that accuracy, and how quickly, and with what pattern, it improves as the target date is approached. In this paper we use glide charts (plots of sequences of root mean squared forecast errors as the target date is approached) to address those questions in the contexts of two models for Arctic sea ice forecasting, the feature-engineered linear regression (FELR) model of Diebold and Göbel (2022), and a new and potentially-superior feature-engineered machine learning (FEML)

* Corresponding author.
E-mail address: goulet_coulombe.philippe@uqam.ca (P. Goulet Coulombe).

[1] All authors have contributed equally.

[2] For example, each summer since 2008 the Sea Ice Prediction Network (SIPN) has sponsored the Sea Ice Outlook (SIO) competition for predicting September average daily Arctic sea ice extent. See https://www.arcus.org/sipn for SIPN, and see https://www.arcus.org/sipn/sea-ice-outlook for SIO. September extent forecasts are produced by many research groups mid-month in June, July, and August, and evaluated once September ends and the outcome is known. Insightful post-season SIO assessments have been produced annually (the most recent is (Bhatt et al., 2022)), and similarly-insightful multi-year retrospective SIO assessments have been produced occasionally (Stroeve et al., 2014; Hamilton and Stroeve, 2016; Hamilton, 2020).

https://doi.org/10.1016/j.eneco.2023.106833
Received 20 June 2022; Received in revised form 13 June 2023; Accepted 19 June 2023
Available online 8 July 2023

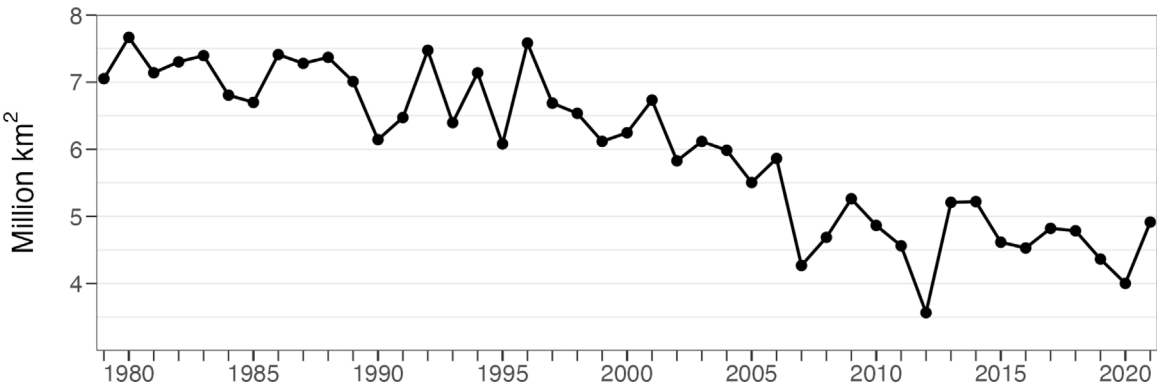


Fig. 1. Observed September Arctic Sea Ice Extent. Notes: We show the evolution of the monthly average of Arctic Sea Ice Extent during September between 1979 and 2021.

model developed in this paper, building on the tree-based "macro random forest" of Goulet Coulombe (2020a).

We proceed as follows. In Section 2 we review the FELR model, display and discuss its glide charts for each month of the year, and compare them to those of a naive (pure trend) benchmark. In Section 3 we introduce the FEML model, display and discuss its glide charts for each month of the year, and compare them to those of a different and more sophisticated benchmark, FELR. Hence FELR appears throughout, but in different roles. It appears first in Section 2 as a candidate model to be compared to a naive benchmark, and then in Section 3 as a more sophisticated benchmark against which a potentially even *more* sophisticated candidate model is compared. We conclude in Section 4.

2. Glide charts for Feature-Engineered Linear Regression (FELR)

Here we review the FELR model of Diebold and Göbel (2022) and display its glide charts for Arctic sea ice forecasting. In particular, we treat FELR as a simple but hopefully-sophisticated model – in the tradition of the "KISS Principle" of forecasting: "Keep it sophisticatedly simple" (Zellner, 1992) – and assess its fixed-target forecasting performance relative to a naive benchmark forecast, a simple linear trend. We do so in part to illustrate the construction and interpretation of glide charts, and in part because we are interested in FELR and the improvements it may deliver relative to more naive models. Later, in Section 3, we turn the tables and use FELR as the benchmark when assessing a more sophisticated non-parametric nonlinear feature-engineered machine learning model.

2.1. Feature-Engineered Linear Regression

To understand FELR, one must understand the real-time fixed-target forecasting exercise in which it is embedded. In our subsequent empirical work, we will consider fixed-target forecasting for a selected target month (the monthly average of daily observations), conditioning on the expanding daily historical sample as the end of the target month is approached, performing 120 daily estimations and making 120 corresponding fixed-target forecasts, starting 120 days before the last day of the target month and continuing to the last day of the target month.[3] Many variations and extensions (e.g., forecasting a particular target day rather than a target monthly average) can be implemented. Although our framework is applicable to fixed-target forecasting of any variable, our subsequent empirical work will focus on Arctic sea ice extent (SIE) and we have specialized the notation below to this particular exercise. Given the importance of seasonality not only in intercepts, but also in trends and dynamics (Diebold and Rudebusch, 2022), we run regressions for each month separately.

Fully general notation gets tedious, so we take a specific example. Consider fixed-target forecasting for September average daily sea ice extent, SIE_9, conditioning on the expanding historical sample as we move from June through the end of September. In FELR, September extent is regressed on an intercept, a linear trend term, and three additional covariates:

$$SIE_9 \rightarrow c,\ Time,\ SIE_{LastMonth},\ SIE_{Last30Days},\ SIE_{Today}, \quad (1)$$

where SIE_9 denotes September average daily extent, "→" denotes "is regressed on", and the rest of the notation is obvious.[4]

As a concrete illustration, and approximately following the SIO forecasting competition (Bhatt et al., 2022), consider the SIE_9 forecasts on four days: 6/10, 7/10, 8/10 and 9/10. Immediately, the 6/10 regression used to produce the June forecast of September is

$$SIE_9 \rightarrow c,\ Time,\ SIE_5,\ SIE_{5/12\_6/10},\ SIE_{6/10},$$

the 7/10 regression used to produce the July forecast of September is

$$SIE_9 \rightarrow c,\ Time,\ SIE_6,\ SIE_{6/11\_7/10},\ SIE_{7/10},$$

the 8/10 regression used to produce the August forecast of September is

$$SIE_9 \rightarrow c,\ Time,\ SIE_7,\ SIE_{7/12\_8/10},\ SIE_{8/10},$$

and the 9/10 regression used to produce the September forecast of September is

$$SIE_9 \rightarrow c,\ Time,\ SIE_8,\ SIE_{8/12\_9/10},\ SIE_{9/10}.$$

Of course the four days above were chosen just as an illustration, conforming approximately with SIO forecast dates. In reality we can produce a forecast on any of the days before the last day of September.

Perhaps surprisingly given their simplicity, the FELR forecasts are quite sophisticated in certain respects of relevance for sea ice forecasting. First, they capture low-frequency linear trend dynamics via conditioning on $Time$. Second, they capture medium-frequency inertial (autoregressive) dynamics around trend by conditioning on $SIE_{LastMonth}$. Finally, they capture high-frequency dynamics by augmenting the conditioning on historical monthly information (via $SIE_{LastMonth}$) with potentially-invaluable recent daily information, via $SIE_{Last30Days}$ and SIE_{Today}.

Empirical results validate such modeling choices, as FELR is a more-than-adequate benchmark, surpassing the SIO median (the median of

[3] We focus on the monthly aggregate rather than raw daily readings, because the monthly aggregate is the object of interest in many climate studies (see Goulet Coulombe and Göbel (2021) and references therein), and also in the SIO annual forecasting competition (Bhatt et al., 2022). Furthermore, raw daily readings are likely to include undesirable high-frequency noise from satellite measurements and post-processing (Diebold et al., 2021).

[4] Diebold and Göbel (2022) use the term "benchmark predictive model" (BPM) rather than FELR.

all submitted forecasts) for September and the three horizons for which the latter is available. This can be seen in the September subplot of Fig. 4. While the linear trend is widely used as a generic reference point, the SIO median, very much like the mean of the Survey of Professional Forecasters in macroeconomics, is a tenacious contender against which to benchmark new approaches (Andersson et al., 2021). The crucial practical advantage of FELR over the SIO median is obviously that we can generate its forecasts for more than three arbitrarily-fixed dates and a single target month, producing direct (rather than iterated) FELR forecasts day-by-day, using model parameter estimates optimized to the remaining predictive horizon, thanks to the trivial simplicity and speed of FELR estimation by linear least-squares regression.[5] We exploit this fact below to make and examine 120 daily fixed-target Arctic sea ice forecasts from June through September.

2.2. Glide charts

We measure forecast performance, and its evolution as the target date is approached, with *RMSFE glide charts* (RGCs). An RGC is simply a plot of the sequence of root mean squared forecast errors (RMSFEs) from the ordered sequence of 120 regressions with the conditioning information expanding as the target date is approached, where $RMSFE = \sqrt{\frac{e'e}{T}}$, for regression residual vector e and sample size T.[6]

We use SIE observational data from 1979 to 2020 from the National Snow and Ice Data Center (NSIDC), which uses the NASA team algorithm to convert microwave brightness readings into ice coverage (Fetterer et al., 2017a).[7] We start with a 4-month lead (120 days). Once the 120 FELR regressions are run, we construct RGCs. RGCs naturally differ across months, so we examine the shapes for each month separately.

In Fig. 2 we show the twelve monthly FELR RGCs, together with the twelve monthly benchmark linear trend RGCs for comparison. First consider the linear trend RGCs, which are flat, for all months and horizons. This is expected, because they capture only extremely low-frequency dynamics and hence take almost no account of evolving conditions as the target date is approached.

Now consider the FELR RGCs, which are very different. Very early on, near date $T - 120$, they are little different from those of the linear trend benchmark, but accuracy eventually improves (so the RGCs drop), achieving perfection by the target date T (so the RGCs are 0). Moreover, the precise glide chart paths differ across months.

Most months show FELR predictability thresholds, meaning that accuracy initially fails to improve as the target date is approached, but then increases progressively once a predictability threshold lead time is crossed. For example, the peak-summer (July) sea ice forecast shows no increase in accuracy until roughly $T - 60$, after which accuracy rapidly improves. This corroborates the results of Day et al. (2014), who find a spring "predictability barrier" for summer pan-Arctic SIE predictions, but contrasts with those of Bushuk et al. (2019).

Interestingly, the predictability thresholds are earliest for the low-ice months of August, September (when Arctic sea ice achieves its minimum), and October. The August threshold is around $T - 90$, the September threshold is evidently around $T - 120$, and the October threshold is even earlier – literally off the chart!

[5] One makes a multi-period "direct" forecast with a horizon-specific multi-period-ahead estimated model. In contrast, one makes a multi-period "iterated" forecast with a one-period-ahead estimated model, iterated forward for the desired number of periods. Direct projections are theoretically superior under model misspecification (which is always the relevant case), because they directly minimize the relevant multi-step predictive loss, as per (Ing, 2003), Theorem 4 and Corollary 3.

[6] We are of course not the first to work with glide charts or similar constructs (whatever the name) for sea ice forecasts, whether in absolute terms or relative to a benchmark. Key recent references include (Chevallier et al., 2013), Day et al. (2014), Hawkins et al. (2016), and Bushuk et al. (2019).

[7] See Appendix A for details.

3. Glide charts for feature-engineered machine learning (FEML)

Although FELR clearly captures salient features in SIE_m at various forecasting horizons, it remains a linear model. Simultaneously, there are ample plausible sources of nonlinear SIE_m dynamics, including tipping points and feedback loops (Maslanik et al., 2007). With the shape of those nonlinearities being unknown, we turn to flexible nonlinear machine learning (ML) approaches to estimate them. A significant roadblock to that enterprise, however, is that completely nonparametric ML methods simply will not work on a sample of size $T \approx 40$. We confront this situation by using a feature-engineered machine learning (FEML) approach, which, as the name suggests, retains the feature engineering that made FELR a successful benchmark, but with substantial generalization. In particular, the ML model we consider is Macro Random Forest (Goulet Coulombe, 2020a), which builds nonlinearities around FELR rather than modeling everything nonparametrically. Hence FEML continues to capture linear signals precisely as with FELR, but it can also capture additional nonlinear signals, while continuing to economize on degrees of freedom.

To make such points more clear, we first review the basics of Macro Random Forest (MRF) and then describe the various FEML specifications that will be used in our subsequent forecasting exercise.

3.1. Feature-engineered machine learning

Here we introduce a flexible tree-based class of nonparametric nonlinear feature-engineered models for fixed-target sea ice forecasting.

3.1.1. Macro random forest

Goulet Coulombe (2020a) proposes a new form of random forest (RF, Breiman 2001) better suited for time series, especially macroeconomic data where the available series are typically of short length. The model is

$$y_t = X_t\beta_t + \epsilon_t$$
$$\beta_t = \mathcal{F}(S_t)$$

where S_t are the state variables governing time variation and $\mathcal{F}$ a forest. S_t can be a large data set, beyond what is included in FELR. X determines the *linear* model that we want to be time-varying. By design, it is preferable that $X \subset S$ be parsimonious and a priori important compared to the larger S. For instance, one can use lags of y_t for X_t when an appreciable degree of persistence is suspected. In this paper, X_t will be the features of FELR.

While an advantage of the method is its potential for interpretation via the generalized time-varying parameters β_t, of greater interest here are its predictive advantages in an environment with scarce data and a strong linear signal. Indeed, it is easy to see that, if $\mathcal{F}$ ends up hardly nonlinear – or seen differently, mostly time-invariant – FEML collapses to FELR. In contrast, a plain RF that learns nothing, collapses to the unconditional mean. Thus, FEML constructs the conditional mean economically by starting with FELR and incorporating nonlinearities (as much as one can afford with $T \approx 40$) around it. In contrast, a plain RF would struggle to capture linear autoregressive signals effectively using hard-thresholding functions (the trees) and would be left with little or no degrees of freedom for "real" nonlinearities. For much more on this point, see Goulet Coulombe (2020a).

The estimation is carried out through a greedy algorithm, which, in its most basic form, is to run

$$\min_{j\in\mathcal{J}^-,\ c\in\mathbb{R}} \left[\min_{\beta_1} \sum_{t\in l_1(j,c)} \left(y_t - X_t\beta_1\right)^2 + \min_{\beta_2} \sum_{t\in l_2(j,c)} \left(y_t - X_t\beta_2\right)^2 \right]. \quad (2)$$

recursively to construct trees. In words, at each potential tree split, we obtain the optimal variable S_j (i.e., the best j out of the random subset of predictors indexes $\mathcal{J}^-$) with which to split our sample, and c, i.e. the value at which we should split j. The resulting outputs j^*

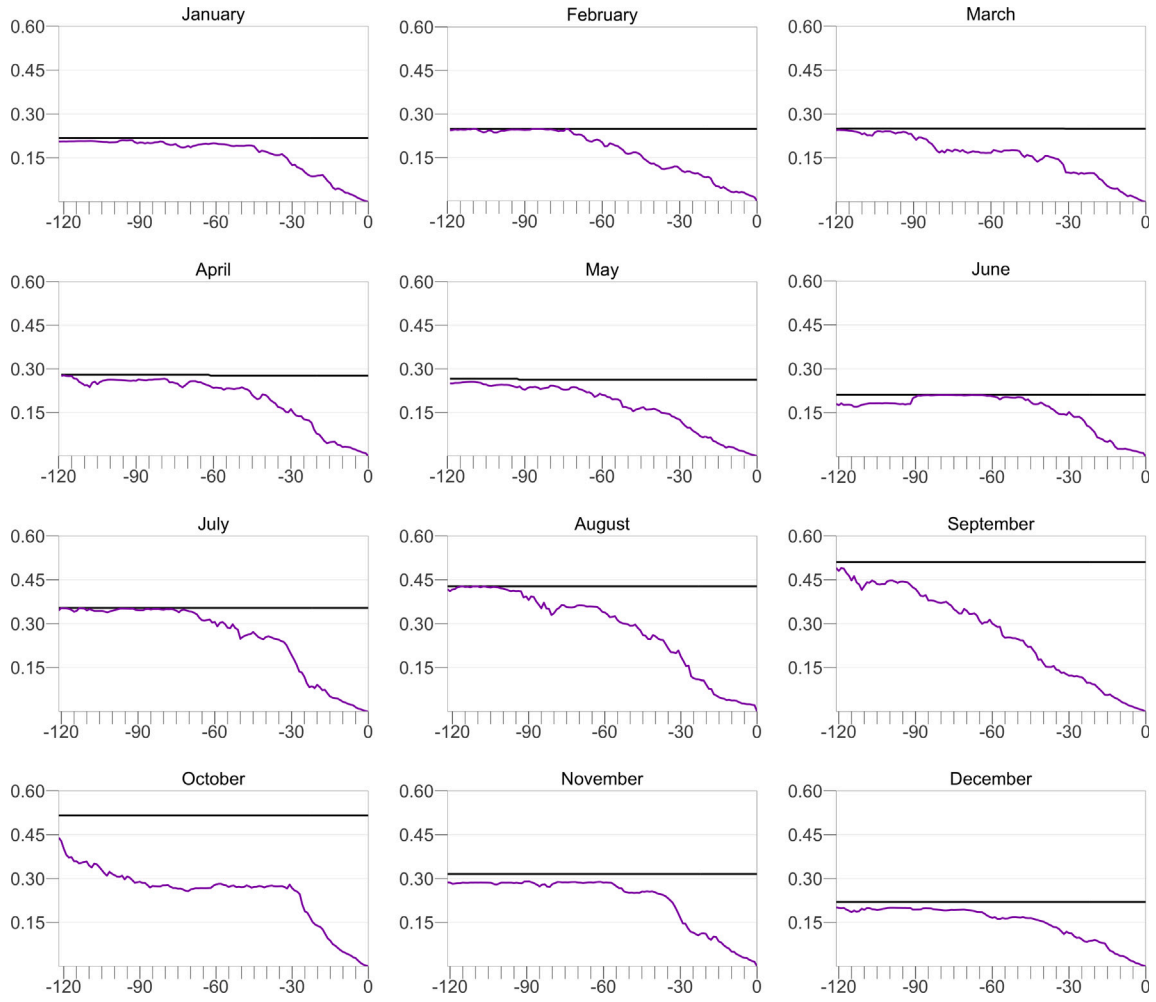


Fig. 2. Glide charts: Feature-engineered linear regression.
Notes: We show RMSFE from linear trend models and feature-engineered linear regression models:
$SIE_m \rightarrow c,\ Time\ \ (black)$
$SIE_m \rightarrow c,\ Time,\ SIE_{LastMonth},\ SIE_{Last30Days},\ SIE_{Today}\ \ (orchid)$,
with the estimation sample expanding over 120 days. The horizontal axes show the number of days until the end of the target month m. In some instances $SIE_{LastMonth} = SIE_{Last30Days}$, so that some models would suffer from perfect multicollinearity. In these cases, we drop $SIE_{Last30Days}$. See text for details.

and c^* are used to split l (the parent "leaf") into two children leaves, l_1 and l_2. Splitting things in halves, and those halves in other small halves eventually leads to obtaining leaves of size 1 (or a small number) yielding β_t, a coefficient at each point in time.

As in RF, the core sources of regularization in MRF are (1) averaging over a diversified ensemble of trees generated by Bagging, and (2) random eligibility of predictors for splits $\mathcal{J}^- \subset \mathcal{J}$.[8] Nonetheless, β_t's (and corresponding predictions) may benefit from additional regularization—this is particularly true of short time series where the potency of Bagging is more limited. Time smoothness is made operational by taking the "rolling-window view" of time-varying parameters. That is, the tree solves many weighted least squares problems including close-by observations. To keep computational demand low, (Goulet Coulombe, 2020a) suggests to use a kernel $w(t;\zeta)$ designed as a symmetric 5-step Olympic podium. Informally, the kernel puts a weight of 1 on observation t, a weight of $\zeta < 1$ on observations $t-1$ and $t+1$ and a weight of ζ^2 on observations $t-2$ and $t+2$. Finally, a small Ridge penalty is added for matrices to invert nicely even in very small leaves, which will be inevitably prevalent in our application. With those additions, (2) turns into the more sophisticated

$$\min_{j\in\mathcal{J}^-,\ c\in\mathbb{R}} \left[\min_{\beta_1} \sum_{t\in l_1^{RW}(j,c)} w(t;\zeta)\left(y_t - X_t\beta_1\right)^2 + \lambda\|\beta_1\|_2 + \min_{\beta_2} \sum_{t\in l_2^{RW}(j,c)} w(t;\zeta)\left(y_t - X_t\beta_2\right)^2 + \lambda\|\beta_2\|_2 \right], \tag{3}$$

where $l_1^*(j,c)$ and $l_2^*(j,c)$ denote the expanded leaves incorporating the aforementioned neighboring observations in time space.

To put things in perspective, a standard RF is a restricted version of MRF where $X_t = \iota$, $\lambda = 0$, $\zeta = 0$ and the block size for Bagging is 1. Said differently, the sole regressor is an intercept, there is no within-leaf shrinkage, and Bagging is carried out as-if we were working with a cross-section. As discussed earlier, by design, MRF will have an edge over RF whenever linear signals included in X_t are strong and the number of training observations (or signal-to-noise ratio) is low. Clearly, all those boxes are checked in this paper's application.

3.1.2. Two specifications

We consider two MRF specifications corresponding to different configurations of S_t and X_t. First, the FEML model has a linear part X_t comprising the very same features of the FELR model, i.e., an

[8] This, and the fact that β_t comes from very small leaves obtained from running (2) recursively, is precisely why we get a different β_t for each t.

intercept (c), a linear time trend ($Time$), SIE of the previous month ($SIE_{LastMonth}$), the average SIE over the last 30 days ($SIE_{Last30Days}$), and today's measurement of the Arctic's sea ice extent (SIE_{Today}). State variables S_t feature a larger set of potentially informative climate variables, akin to Goulet Coulombe and Göbel (2021)'s VARCTIC, designed to proxy the current state of the Arctic. In particular, we include (1) all features entering X_t, (2) daily SIE measurements of the previous 14 days, (3) daily Sea Ice Thickness (SIT) measurements of the latest 14 days available, (4) the average SIT over the latest 30 days of available measurements, (5) lags of monthly measurements of SIE, SIT, CO2 and Air Temperature (AT), and (6) the first five principal components of the feature set described in (1)-(5), which may help in summarizing the key variation in the relatively large S_t.[9,10] Details on the provenance of the various data series appear in Appendix A.

Second, the "Pocket FEML" model has the full S_t of FEML but X_t is a subset of FELR's regressors, namely c, $Time$, and SIE_{Today}. The motivation for a restricted FELR as a linear part is that the size of X_t ultimately reduces the potential depth of trees in the forest for very small datasets. This is due to the fact that the algorithm needs to run small ridge regressions in each leaf, and the larger that regression gets, the larger the minimal leaf size must be to accommodate that operation. In short, it limits the expressivity of the trees by restricting their depth. Thus, the potential benefits of a more condensed X_t is to discard partially redundant information, avoid near-singularity problems in small terminal nodes, and ultimately leave more room for nonlinearities in $\mathcal{F}$. The cost is that, obviously, with respect to the MRF specification, we lose the linear signals from the less noisy $SIE_{LastMonth}$ and $SIE_{Last30Days}$ (although they are included in S_t). While the necessity of those is uncertain prior to 30-day-ahead forecasts, they are mechanically essential for short-run forecasts. This will be clearly visible in the empirical results. Obviously, this is known ex ante and a forecaster can simply switch to FELR or FEML past that threshold.

Regarding tuning parameters, we set values that are a priori more adequate in an environment with very little data. The sampling rate for features in S_t is $\frac{1}{3}$, which is standard. The subsampling rate of data rows is $\frac{9}{10}$ which is rather high and limits the potential for tree-diversification coming from that source. The upside is that it allows for slightly deeper trees to be grown, which is much needed when faced with a small T. ζ is set to 0.25 which reflects the view that we expect little persistence in the underlying β_t's at the yearly frequency. λ is 1 (higher than what is used in typical macroeconomic specifications), and brings helpful regularization when both the data subsampling rate is high and ζ is low. The prior mean for the ridge shrinkage is switched from 0 to values of OLS coefficients, which reflects the view, like in the choice of a higher λ and the specific X_t's, that if there is an improved model to be found, it should not be excessively far from FELR. Our main out-of-sample results are robust to non-trivial deviations in both λ and ζ.

[9] Daily SIT measurements are published at the end of the following month, i.e. on April 25th the time series covers data only through the end of February. The data for March is not released until May 1st. We use data that is publicly available at the time of forecast. Consequently, depending on the exact day at which one is making a prediction, the SIT enters S_t with a lag of one to two months.

[10] The number of lags differs by variable, but all have a common starting point. For example, when making a prediction on January 20th of year t, the monthly lags for SIE, SIT, CO2 and AT start with a measurement for January of year $t-1$, and end with the latest month on which complete information is available. Thus, for SIE this boils down to 12 monthly lags from January of $t-1$ until December of $t-1$. For SIT, we only have information until and including the whole month of November $t-1$. Finally, for CO2, the monthly lags run from January $t-1$ through October $t-1$, and AT enters with monthly estimates for January $t-1$ through September $t-1$.

3.2. Glide charts

Here we display glide charts for our two FEML versions (FEML, Pocket FEML) and compare them to glide charts for our two FELR versions (FELR, Pocket FELR).[11] We also distinguish between in-sample and out-of-sample versions.

3.2.1. An in-sample analysis

In Fig. 3 we show day-by-day in-sample RMSFEs of selected FELR and FEML models for each month. Calculation of RMSFE requires training set residuals. While those are perfectly fine to use for linear regressions, they are not for random forest-based models. It is customary that successful (as per test set performance) RFs vastly overfit the training data, nearly eliminating the presence of residuals (Goulet Coulombe, 2020b) – i.e., a form of "benign overfitting". Consequently, it is typical to rely on the so-called out-of-bag error to internally evaluate the goodness-of-fit from such models (Breiman, 2001). We do so using block subsampling as in Goulet Coulombe (2020a) which is more adequate in the context of time series data. Here, we set the block size to two years.

Mechanically, we observe that among the FELRs, the model with the smallest degrees of freedom (FELR, with 5 parameters) is the lower envelope. It is noted that the differences between FELR and Pocket FELR are often small, except for the longer horizons of spring and summer months. This indicates that the cost of forgoing certain regressors in Pocket FEML's linear part may not be too wise a choice in the forthcoming out-of-sample evaluation. In this in-sample evaluation, FEML RMSFEs are higher than those of FELRs in almost every instance. However, it is worth remembering that RMSFEs are computed differently (by necessity) and that FEMLs' calculations account for degrees of freedom while FELRs' do not. To provide an apples-to-apples comparison of the competing FE approaches, we switch to a uniform recursive out-of-sample evaluation metric.

3.2.2. A pseudo-out-of-sample analysis of the last decade

Glide charts need not necessarily be used in conjunction with RMSFE based on in-sample residuals or variants of them. In fact, any loss can be used. In Fig. 4, we remain within the realm of squared errors, but those are computed from a recursive expanding-window pseudo-out-of-sample experiment. This sort of exercise is standard in the modern macroeconomic forecasting literature comparing econometric and machine learning models (Goulet Coulombe et al., 2022). The choice of the 2012–2021 window for the "test" set is inspired by Andersson et al. (2021).[12] Given data limitations, it is a fair balance between avoiding training models on too small of a sample size and calculating RMSFE based on too few out-of-sample errors. Models are re-estimated every year to leverage the gradually incoming new data points.

There are benefits and costs to this alternative evaluation setup. The obvious cost is that the test set RMSFE is the average of 10 errors rather than 40 (as considered in the previous section for the in-sample analysis), which inevitably increases the variance of the evaluation metric. The benefits are threefold. First, it is not unthinkable that the magnitude of the last 10 years' forecasting errors is more informative about the near future than that of those in the 1980s and 1990s. Second, semi-flexible trend models (like FELR) will have a built-in advantage for in-sample evaluation over what prevails when one uses such models to *really* forecast next year's SIE. The reason for this is that in-sample

[11] In addition, to disentangle whether Pocket FEML's performance differential comes from nonlinearities versus a sparser inherent linear equation, we also include Pocket FELR in our set of benchmarks.

[12] They conduct an evaluation of SIE predictability for their convolution neural network trained directly on satellite imagery data. The benchmarks they consider are a climate model and a linear time-trend model.

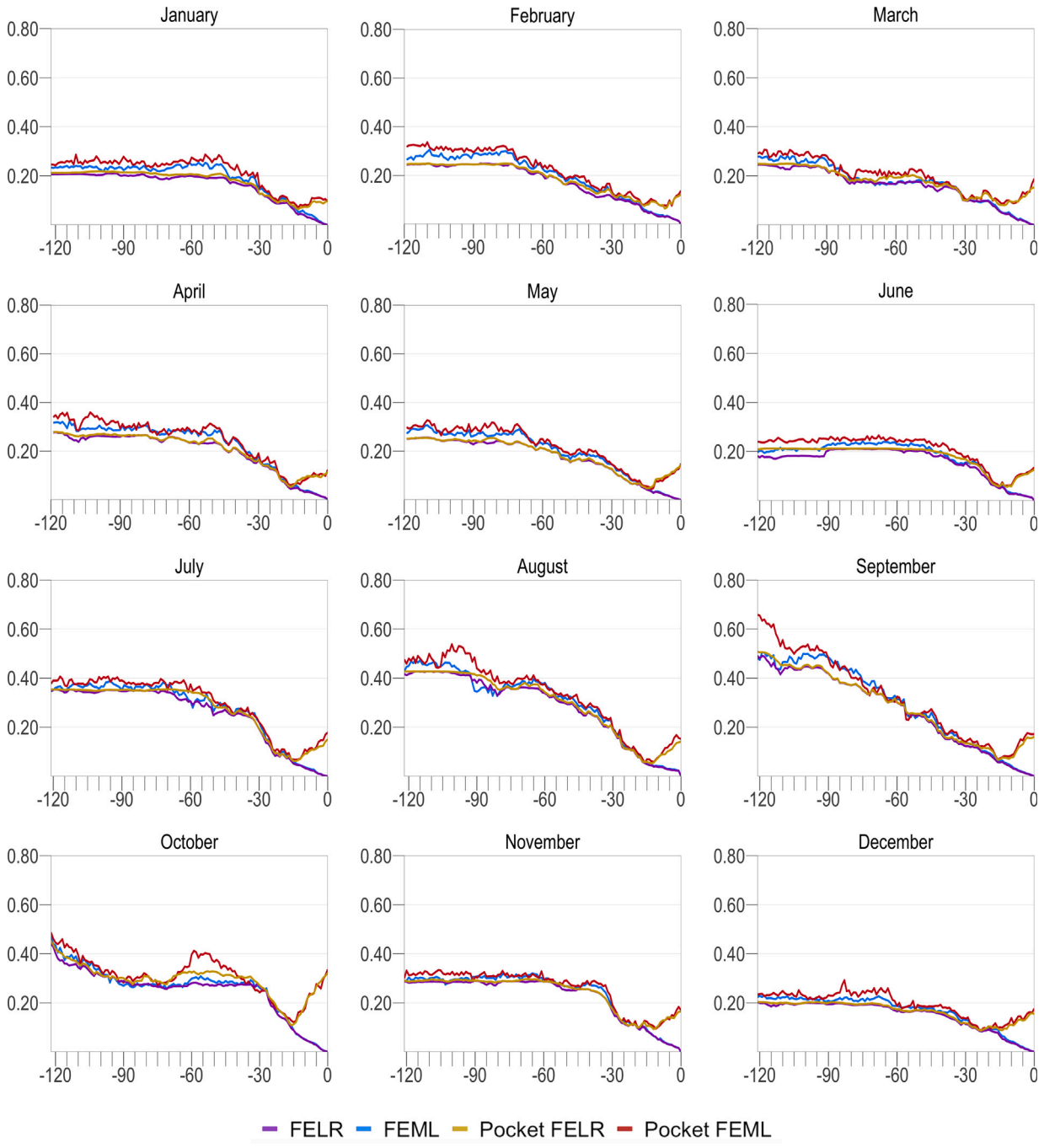


Fig. 3. Glide charts: FELR and FEML.
Notes: We show **in-sample** RMSFE glide charts for several FELR and FEML models. The estimation period is 1979–2020. The horizontal axes show the number of days until the end of the target month. In some instances $SIE_{LastMonth} = SIE_{Last30Days}$, so some models would suffer from perfect multicollinearity. In such cases, we drop $SIE_{Last30Days}$. See text for details.

evaluation uses a residual at time t from a model that is trained on both $t-1$ and $t+1$. Information from $t+1$ is extremely useful *when estimating the parameters* of a time trend, but such information about $t+1$ is not available when one is truly forecasting t from $t-1$. The recursive evaluation addresses this potential bias by mimicking directly the reality of forecasting every year using a model estimated only on available data at that particular point in time. Lastly, an advantage of recursive estimation in our setting is that OLS-based and RF-based models are now evaluated using an identical metric and differences between performances cannot be attributed to various choices on how to account for degrees of freedom (like setting up the out-of-bag metric).

In this out-of-sample evaluation, a couple of observations are worth mentioning. First, FELR and its pocket counterpart stand out as solid benchmarks, by routinely yielding the smallest RMSFE, which is especially clear for longer horizons of early summer months. Given the small estimation sample limitations, it is not entirely surprising that FEML's reductions in RMSFE are limited in size. There are, however, notable and important exceptions. The first is Pocket FEML's performance from 90 to 45 days ahead for September. Needless to say, if there are any SIE forecasts of superior interest, it is exactly those lead-times prior to the end of September (hence period of the annual SIO forecasting competition). Improvements of this particular FEML

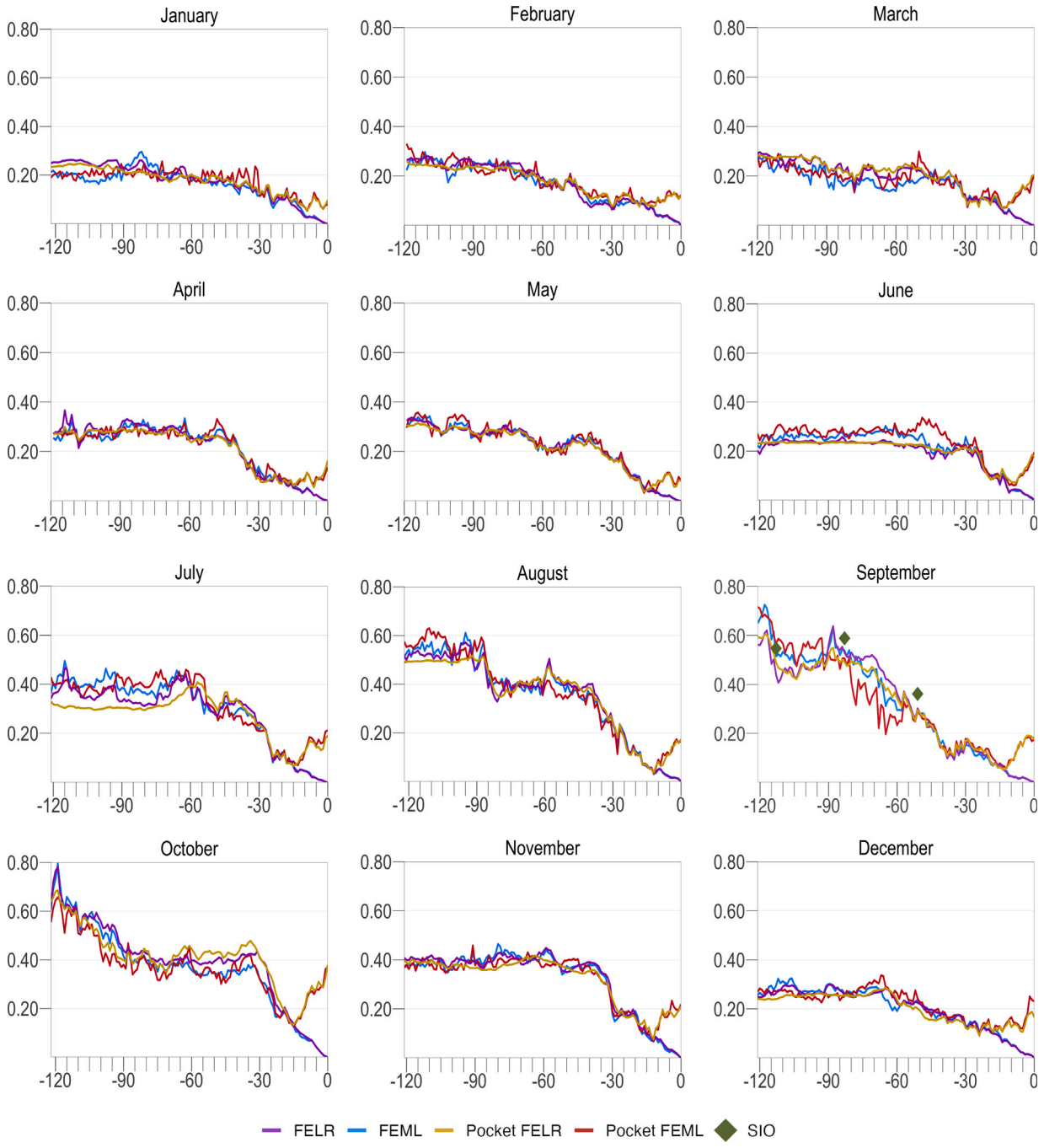


Fig. 4. Out-of-Sample glide charts: FELR and FEML average Over 2012–2021.
Notes: We show **out-of-sample** RMSFE glide charts. We make forecasts each day through the end of the target month in each year from 2012–2021. We then plot the square root of the 10-year average of each day's squared forecast error. The horizontal axes show the number of days until the end of the target month. In some instances $SIE_{LastMonth} = SIE_{Last30Days}$, so some models would suffer from perfect multicollinearity. In such cases, we drop $SIE_{Last30Days}$. See text for details.

over FELRs are over 0.1×10^6 km^2 for the whole period. For October, the dominance of nonlinear models, albeit quantitatively smaller, is present for almost all horizons up to 15 days ahead. Finally, FEMLs (with FEML leading among them) also outperform FELRs for the vast majority of horizons for March—sometimes offering reduction up to 50% in RMSFE.

In Fig. 5, we report the fraction of days for which any FEML in Fig. 4 offers the lowest RMSFEs. This helps summarizing and synthesizing the abundant information in glide charts. It is clear that October, March, and to a lesser extent, January, are all months where gains (albeit small for certain horizons) are generalized over the whole 120 days. Their respective shares of optimal forecasts are above 90%, 80%, and 70% respectively. In contrast, September reductions in RMSFE are substantial in size but are localized within a specific forecasting range. Accordingly, the fraction of days in which any FEML outperforms FELR for September is around 50%. Many months exhibit fractions in such a range, but unlike September, they are typically due to FEML and FELR forecasts being roughly similar. Overall, the early summer months June and July are better predicted using FELRs, and this is mostly attributable to long-range forecasts made in the first months of the melting phase. As can be seen in the bottom quadrants of Fig. 5, the opposite can be said for mid-range horizons where FEMLs are very frequently the best option for many months (excluding June, July and August).

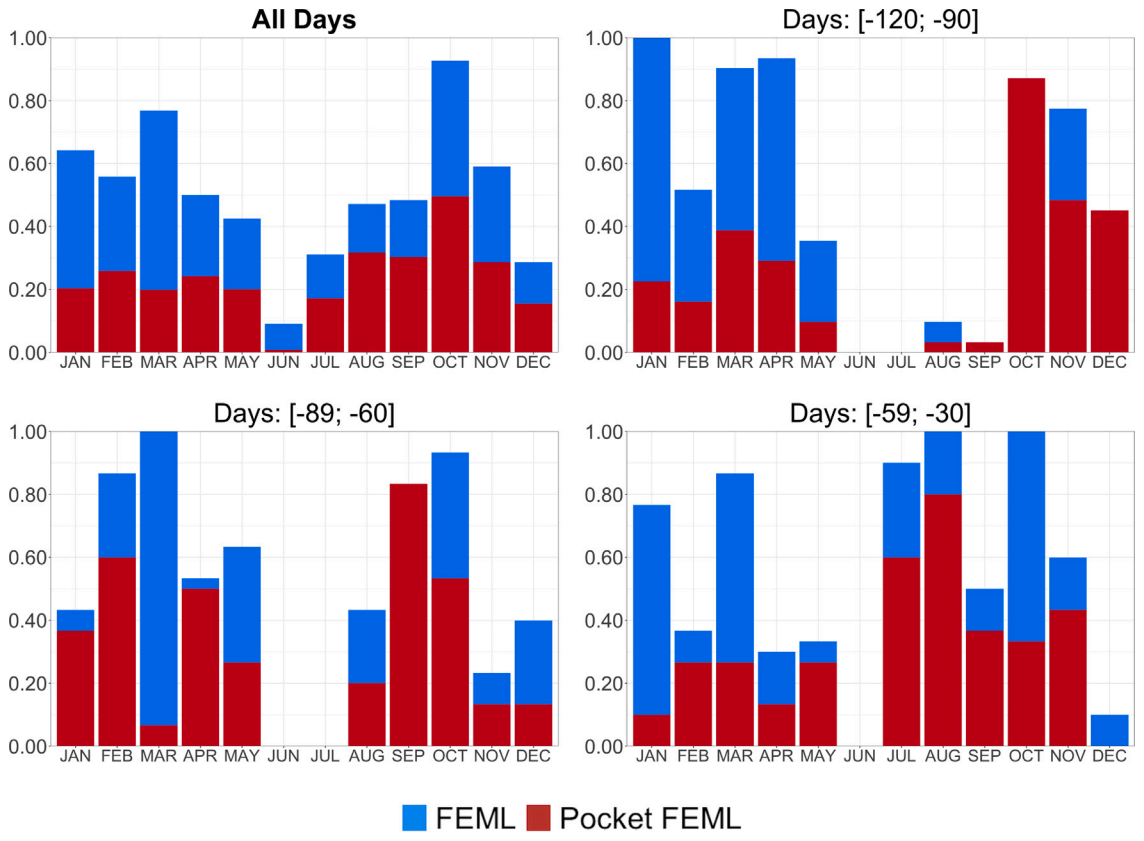


Fig. 5. Fraction of days in which any FEML outperforms any other FELR or FEML Years 2012–2021.
Notes: We show the fraction of days for which each FEML model showed the lowest root-mean-squared forecast errors of out-of-sample predictions between 2012–2021. That is, in each year from 2012–2021 we calculate the squared point-forecast error made on each day through the end of the target month. We calculate the square-root of the 10-year average squared forecast error for each day (see Fig. 4). We then plot the fraction of days, for which each FEML model achieved the lowest 10-year RMSFE. The set of all models includes "FELR", "Pocket FELR", "FEML", and "Pocket FEML". The horizontal axes show the number of days until the end of the target month m. See text for details. Notice that in some instances $SIE_{LastMonth} = SIE_{Last30Days}$ and some models would suffer from perfect multicollinearity. In these cases, $SIE_{Last30Days}$ is dropped. This causes the red and green line to sometimes coincide.

The fact that FEMLs offer clear gains for September, October, *as well as March* forecasts, and much less so for other months suggests that nonlinearities (and an expanded data set) are particularly beneficial for detecting turning points in the annual SIE cycle, that is, when SIE stops expanding or stops retracting. A working hypothesis is that nonlinearities and additional information helps in avoiding either too low or too large SIE predictions around the trough based on slowly evolving physical limits of the seasonal component. In the case of September and October, this could be due to FEML's moderate downward pressures on the prediction from very low readings of SIE_{Today} and $SIE_{LastMonth}$ during early Arctic summer to account for the fact that as ice melts, perhaps more than previous summers, the weighting of multi-year thicker and older ice increases, ultimately slowing the melting process in late summer (Maslanik et al., 2007). Another potential source of nonlinearity, now in favor of accelerated melting beyond what a linear dynamic relationship suggests, is the presence of feedback loops, like the ice-albedo effect, which can manifest even within short time spans (Goulet Coulombe and Göbel, 2021).

Lastly, it can be informative to look at the raw series and corresponding forecasts themselves for the key month of September. Fig. 6 shows three horizons where disagreement between linear and nonlinear models can be substantial (June 14th, July 25th, August 13th). We see FELR $\succ$ Pocket-FEML in June is due to the latter being overly pessimistic in the first half of the out-of-sample period. Disagreement inevitably shrinks in July as the target date approaches. Nonetheless, Pocket FEML clearly gets the upper hand in the early 2010s by better capturing the large deviations from trend starting in 2012 (the lowest SIE on record). Finally, forecasts converge to near-identical values by mid-August.

3.2.3. *Which model(s) to use and when*

As mentioned earlier, *Pocket* models are mechanically handicapped for horizons less than 30 days by excluding the slowly accumulating September data. Fortunately, the glide chart's vocation is to recommend a model to use, and that recommendation may depend on the horizons of interest. In the case of September, the outcome is clear: one should use FELR or Pocket FELR up to 90 days ahead, then switch to Pocket FEML for the next 60 days, and then revert back to FELR for the remaining 30 (short-run) horizons. Glide charts prove to be a particularly useful analytical tool in this exercise given that the optimal model choice for various months is clearly horizon-dependent.

In unreported results, we considered FEML ($S_t = X_t$), a MRF with the linear part X_t still being FELR, but with the set of "forest" variables S_t being restricted to only include the elements in X_t. Naturally, this helps in gauging how much of FEML gains/losses are attributable to the use of a larger information set vs plain nonlinearities. In the overwhelming majority of cases, FEML supplants or performs equally well as FEML ($S_t = X_t$). This suggests that focused nonlinearities *and* an expanded data set can provide the largest gains over FELR. Thus, in line with (Goulet Coulombe, 2020a)'s observations in macroeconomic forecasting applications, a larger S_t is almost always preferable to a restricted one. Moreover, as noted in Goulet Coulombe (2020b), a larger S_t spurs diversification of the underlying trees which helps to keep overfitting in check. Given the short length of our time series, potential for tree diversification is more easily obtained from feature randomization than from bagging.

In sum, FEMLs can provide timely forecasting improvements over FELRs for important months in the SIE annual cycle. Given the data

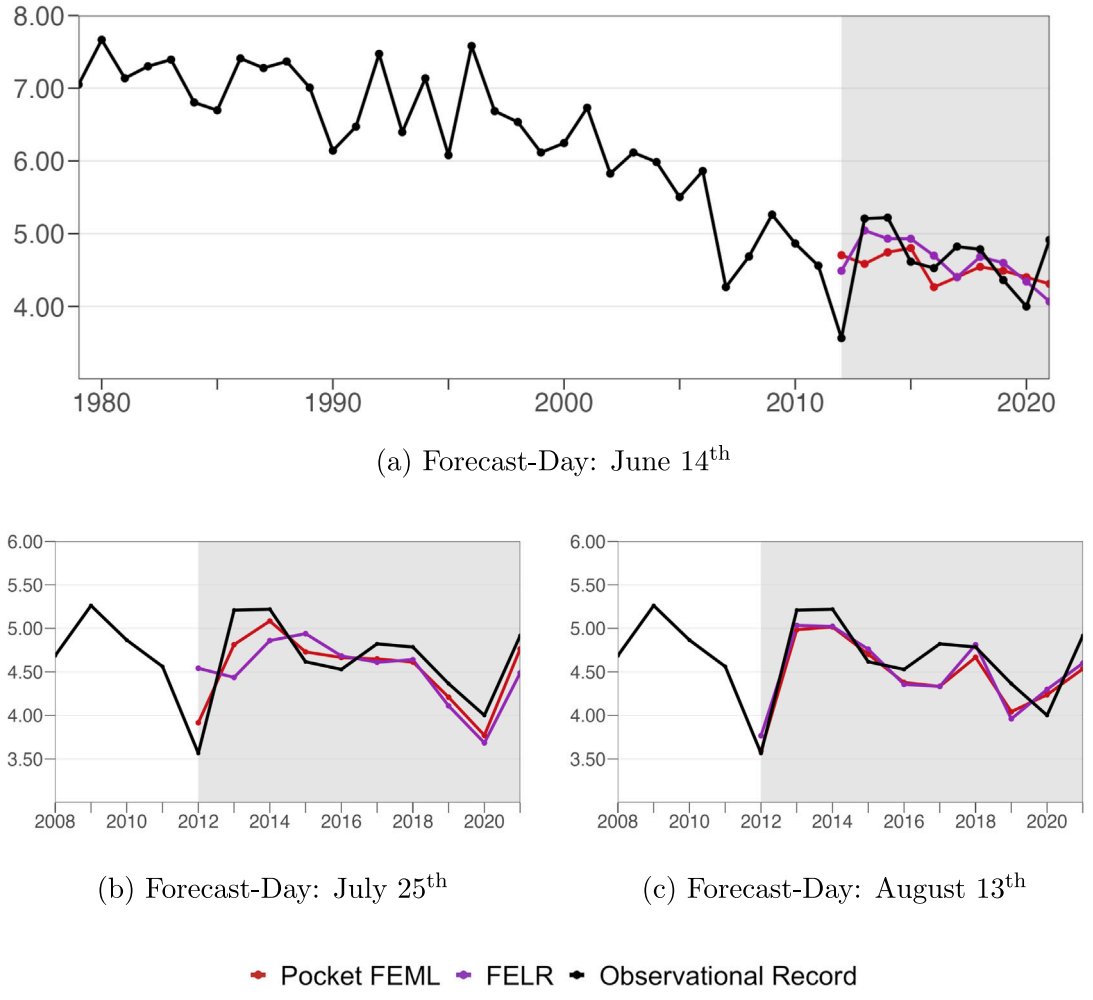


Fig. 6. Annual out-of-sample forecasts on different days.
Notes: We show out-of-sample forecasts for September SIE for the years 2012–2021. The black line is the realized SIE. The scale of the Y-axis is in 10^6 km^2.

limitation, these are not extremely large and are not observed for every horizon, even in successful months. In such a context, glide charts are particularly useful to provide guidance on which feature-engineered model to use and when. Our results unequivocally indicate that FELRs and FEMLs are more adequate benchmarks for out-of-sample predictive accuracy than the oversimplistic linear trend model — which is nonetheless widely used for such purposes (Andersson et al., 2021).

4. Concluding remarks and directions for future research

We have used glide charts – plots of sequences of root mean squared forecast errors as the target date is approached – to evaluate and compare fixed-target forecasts of Arctic sea ice. We first used them to evaluate the feature-engineered linear regression (FELR) forecasts of Diebold and Göbel (2022), and to compare them to naive pure-trend benchmark forecasts. Then we introduced a much more sophisticated feature-engineered *machine learning* (FEML) model, and we used glide charts to evaluate its forecasts and compare them to a FELR benchmark. Our substantive results include the frequent appearance of predictability thresholds, which differ across months, meaning that accuracy initially fails to improve as the target date is approached but then increases progressively once a threshold lead time is crossed. We also compared FELR and FEML, finding that FEML can improve on FEML for turning point months in the annual SIE cycle, namely September, October, and March. Those gains are particularly evident for forecasts made 90 to 30 days before the target date.

In addition, we have built a website that expands on the analysis of this paper, providing weekly updates of forecasts for target date September 2022.[13] The forecasts are based on FELR models, FEML models, and the VARCTIC model of Goulet Coulombe and Göbel (2021). The user can explore the 2022 forecasts and those of previous years through a series of interactive plots. Among other things, the site features a glide chart of the key models as well as a continuously-updated rolling history of 2022 point forecasts and associated prediction intervals. This provides publicly available real-time SIE predictions from four competitive statistical/machine learning models. It therefore complements the Sea Ice Outlook, which is of much larger scope in terms of included models, but which publishes the results of the survey only on a monthly basis and with a lag of two to three weeks.

Several directions for future research are apparent, all of which are related to this paper's use of glide charts for comparing different fixed-target SIE forecasting models, and the related idea that a "better" model should have a "better" glide chart. First, although in this paper we focused exclusively on comparing glide charts of statistical/econometric sea ice forecasting models, one could also (a) include glide charts of structural global climate models (GCMs) (e.g., "How well does a particular GCM's glide chart match the FEML glide chart?"), or (b) use glide charts to help calibrate/estimate GCMs (e.g., "For a particular GCM, what parameter configuration minimizes the divergence between the GCM glide chart and the FEML glide chart?").

[13] See https://chairemacro.esg.uqam.ca/arctic-sea-ice-forecasting/?lang=en.

Second, one could consider glide-chart loss functions, or accuracy measures, other than the ubiquitous quadratic loss underlying RMSFE. In particular, one may want to entertain asymmetric loss functions. Consider, for example, a firm contemplating in June whether to attempt a September trans-Arctic shipment, using September fixed-target sea ice forecasts to guide the decision, and consider positive vs. negative forecast errors:

1. Positive errors (realized September ice greater than forecast): the overly-optimistic forecasts may produce a decision to undertake shipping, which may be regretted as the shipping will be more risky and costly than expected, or even impossible.
2. Negative errors (realized September ice less than forecast): the overly-pessimistic forecasts may produce a decision not to undertake shipping, which may be regretted as business is lost unnecessarily.

Both positive and negative errors are of course costly, but there is no reason why the loss associated with a given positive error should necessarily match that of a negative error of the same absolute magnitude. Asymmetric loss functions capture such effects.

Appendix A. Data

Daily SIE data are from Sea Ice Index, Version 3 (Fetterer et al., 2017b) provided by the National Snow and Ice Data Center (NSIDC).[14] Until August 1986, data are reported only every other day. For model estimation we fill missing $SIE_{Today,t}$ observations with the average of the two adjacent days, $SIE_{Yesterday,t}$ and $SIE_{Tomorrow,t}$.

Daily SIT data are from PIOMAS provided by the Polar Science Center.[15] Data for month m are not known until the end of month $m+1$. Hence, on any day prior to the end of month $m+1$, information on SIT is only available through the end of month $m-1$. This results in a one-to two-month lag.

Daily AT data are based on Rohde and Hausfather (2020).[16] The monthly measurements are reported as anomalies relative to the January 1951–December 1980 average.

Monthly CO2 concentration data are from Mauna Loa, provided by the NOAA Global Monitoring Laboratory.[17] The data for month m are made available during the first days of month $m+1$.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eneco.2023.106833.

References

Andersson, T.R., Hosking, J.S., Pérez-Ortiz, M., Paige, B., Elliott, A., Russell, C., Law, S., Jones, D.C., Wilkinson, J., Phillips, T., et al., 2021. Seasonal Arctic sea ice forecasting with probabilistic deep learning. Nature Commun. 12 (1), 1–12.

Bekkers, E., Francois, J.F., Rojas-Romagosa, H., 2016. Melting ice caps and the economic impact of opening the northern sea route. Econom. J. 128, 1095–1127.

Bhatt, U.S., Meier, W., Blanchard-Wrigglesworth, E., Massonnet, F., Goessling, H., V., Ludwig, Bieniek, P., Eicken, H., Fisher, M., Hamilton, L.C., Little, J., Overland, J.E., Serreze, M., Steele, M., Stroeve, J., Walsh, J., Wang, M., Wiggins, H.V., 2022. In: Turner-Bogren, B., Wiggins, H.V. (Eds.), Sea Ice Outlook: 2022 Post Season Report. URL https://www.arcus.org/sipn/sea-ice-outlook/2022/postseason.

Breiman, Leo, 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Bushuk, M., Msadek, R., Winton, M., Vecchi, G., Yang, X., Rosati, A., Gudgel, R., 2019. Regional Arctic sea–Ice prediction: Potential versus operational seasonal forecast skill. Clim. Dynam. 52 (5), 2721–2743.

Chevallier, M., Salas y Mélia, D., Voldoire, A., Déqué, M., Garric, G., 2013. Seasonal forecasts of the pan-Arctic sea ice extent using a GCM-based seasonal prediction system. J. Clim. 26 (16), 6092–6104.

Day, J.J., Tietsche, S., Hawkins, E., 2014. Pan-Arctic and regional sea ice predictability: Initialization month dependence. J. Clim. 27 (12), 4371–4390.

Diebold, F.X., Göbel, M., 2022. A benchmark model for fixed-target Arctic sea ice forecasting. Econom. Lett. 215, 110478.

Diebold, F.X., Göbel, M., Goulet Coulombe, P., Rudebusch, G.D., Zhang, B., 2021. Optimal combination of Arctic sea ice extent measures: A dynamic factor modeling approach. Int. J. Forecast. 37 (4), 1509–1519.

Diebold, F.X., Rudebusch, G.D., 2022. Probability assessments of an ice-free Arctic: Comparing statistical and climate model projections. J. Econometrics 231, 520–534.

Ebinger, C.K., Zambetakis, E., 2009. The geopolitics of Arctic melt. Int. Aff. 85 (6), 1215–1232.

Fetterer, F., Knowles, K., Meier, W., Savoie, M., Windnagel, A.K., 2017a. Sea ice index. NSIDC: National Snow and Ice Data Center, Boulder, Colorado, USA, Version 3. https://doi.org/10.7265/N5K072F8. Dataset ID G02135, updated daily.

Fetterer, F., Knowles, K., Meier, W.N., Savoie, M., Windnagel, A.K., 2017b. Sea ice index, Version 3. https://doi.org/10.7265/N5K072F8.

Goulet Coulombe, P., 2020a. The macroeconomy as a random forest. Available at SSRN 3633110.

Goulet Coulombe, P., 2020b. To bag is to prune. arXiv e-prints arXiv–2008.

Goulet Coulombe, P., Göbel, M., 2021. Arctic amplification of anthropogenic forcing: A vector autoregressive analysis. J. Clim. 34, 5523–5541.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., Surprenant, S., 2022. How is machine learning useful for macroeconomic forecasting? J. Appl. Econometrics 37 (5), 920–964.

Hamilton, L., 2020. 1000 Predictions: What's new and what's old in a retrospective analysis of the sea ice outlook, 2008–2020. In: Presentation at American Geophysical Union Annual Meeting.

Hamilton, L.C., Stroeve, J., 2016. 400 Predictions: the SEARCH sea ice outlook 2008–2015. Polar Geogr. 39 (4), 274–287.

Hawkins, E., Tietsche, S., Day, J.J., Melia, N., Haines, K., Keeley, S., 2016. Aspects of designing and evaluating seasonal-to-interannual Arctic sea-ice prediction systems. Q. J. R. Meteorol. Soc. 142 (695), 672–683.

Ing, C.-K., 2003. Multistep prediction in autoregressive processes. Econom. Theory 19 (2), 254–279.

Maslanik, J.A., Fowler, C., Stroeve, J., Drobot, S., Zwally, J., Yi, D., Emery, W., 2007. A younger, thinner Arctic ice cover: Increased potential for rapid, extensive sea-ice loss. Geophys. Res. Lett. 34 (24).

Petrick, S., Riemann-Campe, K., Hoog, S., Growitsch, C., Schwind, H., Gerdes, R., Rehdanz, K., 2017. Climate change, future Arctic sea ice, and the competitiveness of European Arctic offshore oil and gas production on world markets. Ambio 46 (3), 410–422.

Rohde, R.A., Hausfather, Z., 2020. The Berkeley earth land/ocean temperature record. Earth Syst. Sci. Data 12, 3469–3479.

Shalina, E.V., Johannessen, O.M., Sandven, S., 2020. Changes in Arctic sea ice cover in the twentieth and twenty-first centuries. In: Johannessen, O.M., Bohylev, L.P., Shalina, E.V., Sandven, S. (Eds.), Sea Ice in the Arctic: Past Present and Future. Springer Nature, pp. 93–166.

Stroeve, J., Hamilton, L.C., Bitz, C.M., Blanchard-Wrigglesworth, E., 2014. Predicting september sea ice: Ensemble skill of the SEARCH sea ice outlook 2008–2013. Geophys. Res. Lett. 41 (7), 2411–2418.

Zellner, A., 1992. Statistics, science and public policy. J. Amer. Statist. Assoc. 87, 1–6.

[14] See https://doi.org/10.7265/N5K072F8.

[15] See http://psc.apl.uw.edu/wordpress/wp-content/uploads/schweiger/ice_volume/PIOMAS.thick.daily.1979.2022.Current.v2.1.dat.gz.

[16] See http://berkeleyearth.lbl.gov/auto/Global/Land_and_Ocean_complete.txt.

[17] See the "deseasonalized" column in https://gml.noaa.gov/webdata/ccgg/trends/co2/co2_mm_mlo.csv.