

Random Analog Prediction of Arctic Sea Ice Extent: a Benchmark for Seasonal Models

Faiq Raees

New York University

Francesco Paparella

`francesco.paparella@nyu.edu`

New York University Abu Dhabi

Article

Keywords:

Posted Date: October 27th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-7868093/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Random Analog Prediction of Arctic Sea Ice Extent: a Benchmark for Seasonal Models.

Faiq Raees^{1,2,3,*} and Francesco Paparella^{2,4,**}

¹Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, USA.

²Division of Science, New York University Abu Dhabi, UAE

³Center for Stability, Instability, and Turbulence, New York University Abu Dhabi, UAE

⁴Mubadala Arabian Center for Climate and Environmental Sciences, New York University Abu Dhabi, UAE

*fr872@nyu.edu

**francesco.paparella@nyu.edu

ABSTRACT

We develop a Random Analogue Predictor (RAP) algorithm for the forecasting of Arctic Sea Ice Extent (SIE) on seasonal timescales. This is a stochastic variant of the celebrated method of the analogues that only uses the historical SIE record to produce ensemble forecasts. When comparing the observations with the most representative forecast of the ensemble (as identified through the band–depth, a centrality measure for functional data) the algorithm shows negligible bias and RMSE no larger than $0.6 \cdot 10^6$ km². We argue that simplicity, interpretability, independence on physical hypothesis, and the ability to attach an uncertainty estimate to its own forecasts, should make RAP the benchmark of choice for physics–based and AI–based models alike.

Introduction

Edward N. Lorenz distinguished two related, but distinct kinds of forecast. Those of the first kind aim at approximating as closely as possible what will actually occur. Those of the second kind aim at generating data that will match some statistics of what will actually occur, while allowing for much difference in the fine details^{1,2}. Weather forecast typically yearn to fall within the first kind (e.g. determining where a hurricane will be three days from now), while climate projections are obvious examples of forecasts of the second kind (e.g. quantifying the average number of hurricanes making landfall some decades into the future).

Long–range and seasonal weather forecasts occupy a non–enviable middle position, where there is the theoretical possibility (and the practical need) to foresee the actual time evolution of some features (typically large–scale temperature and pressure patterns) but the forecast range clearly exceeds the predictability time of most other features³. Arctic sea–ice extent forecast falls in this category: the characteristic time scales of sea–ice extent changes over the whole Arctic ocean are the seasonal and longer ones, which lends credibility to attempts at pinpointing several months in advance quantities such as the September minimum sea–ice extent. However, when this sort of forecasts is performed over seasonal timescales by means of a coupled atmosphere–ocean–ice model, the details of short–lived, fast–moving weather and ocean structures shown by the model inevitably differ from the observed ones. One can then only hope that the phenomenon of interest (the time evolution of the sea–ice extent) is not overly affected by any individual short–lived event which is unpredictable months in advance (e.g. an ice–churning storm) and that the seasonal–scale dynamics remains satisfactorily reproduced by the model, even if the small–scale details are different. The theoretical underpinnings of this hope have long been explored^{4–7} and we shall not delve in them here.

Having to accept the existence of unpredictable aspects in a phenomenon that one would ideally like to model in a deterministic way raises the question of how to evaluate the results of a forecast. Are the discrepancies between the observations and the results of any specific model only due to those unpredictable aspects or are they due to shortcomings of the model? In other words, how can one decide if there are still margins of improvement in model making or if the state of the art is already close to the theoretical best that can possibly be attained?

While an exhaustive, rigorous answer to the above question may not be easy to obtain, a practical approach consists in comparing a proposed model against a dumb benchmark. We would call a benchmark "dumb" if it consists of an oversimplified model that does not exploit all the physics, observations and computing power that one could possibly include in a full–fledged model, but that still retains a significant forecasting skill. Thus, beating the benchmark would not be a significant merit for a full–fledged model, but not beating it might underscore serious shortcomings of the model under test. In the early days of weather forecasts the benchmark of choice was persistence (whose algorithm is simply: "*n* days from now the weather will be the same as today"). In spite of its simplicity and *dumbness*, persistence has been a useful and challenging benchmark for

short-range and medium-range numerical weather forecasts^{8–11}.

For seasonal forecast, one could use persistence on an annual basis (e.g. "the September minimum sea-ice extent this year will be the same as the past year") but, for forecasts looking only one or two months ahead, backtracking to the previous year in order to use persistence may result in a benchmark *too* dumb. More importantly, persistence forecast, in its standard form, does not give any forecast of the forecast skill. This is a quantification of the margin of uncertainty attached to the forecast itself, and is a necessary information for verifying if a model's skill is actually better than the benchmark¹². For short-range weather forecasts, a statistical analysis over a large number of past cases may give useful indications in this sense, but gathering the necessary statistics for seasonal-scale problems might be unfeasible. For these problems, a satisfactory benchmark should also yield a range of possibilities, much as an ensemble forecast would do, thus quantifying, albeit in a rudimentary and hopefully improvable way, what may be the predictability of a single event. In short, a desirable benchmark for sea-ice extent forecasts should have much in common with a forecast of the second kind.

One of the first methods that were used to attempt seasonal-scale forecasts was that of analog forecasting^{11,13}. In a nutshell, the method consists of searching for past instances where the state of the system to be predicted happened to be as close as possible to the currently observed state. Those are called *analogues* of the current state. The forecast would then consist in assuming that the current state will change in time in the same way as an analogue state. If several analogues can be found, then the method yields different distinct forecasts whose spread may be used to quantify the predictability of the current state¹⁴. Distinct flavors of the method exist, differing in the details of the definition of state of the system, and similarity among states. All of them face the practical difficulty that the available observational record may not be extended enough back in time as to contain a satisfactory analogue to the currently observed state^{11,15}.

When the state of a system at a given time is represented, or summarized, by a single number, such as the sea-ice extent, then the dynamics of the system is represented by a time series. The method of analog forecasting can be applied to time series by using the *delay coordinate embedding* technique, first introduced 1980^{16,17} and then studied extensively^{18–26}. This technique consists in embedding the time series in a M -dimensional Euclidean vector space, by associating individual observations to points of the space (see 'Methods'). Points that are nearby in the embedding space are considered to be analog of each other, in the sense that their future evolution is expected to be similar. For low-dimensional deterministic dynamical systems this expectation is well grounded in rigorous theorems^{16,27} which have been partially extended to more realistic cases of noisy data or stochastic processes^{23,28}.

A forecasting technique that is well adapted to geophysical and astrophysical time series is RAP (random analogue prediction), and consists in a random choice, weighted by distance, of the analog states in the embedding space. Due to the random nature of the forecast, repeated runs yield slightly different results. Thus RAP is inherently a forecast of the second kind: it aims at reproducing the statistical properties of the dynamics, rather than to perfectly match the observations. On the other hand, inasmuch the underlying time series is representative of a deterministic dynamics, that is preserved by RAP².

Thus, RAP appears to be an ideal candidate as a *dumb* benchmark for seasonal-scale forecasts of aggregate data, such as sea-ice extent: it is simple to implement, does not rely on any specific physical hypothesis, parameterization or approximation (it is purely data-driven), and its stochastic nature naturally leads to ensembles of forecasts, that may be used to quantify (albeit imperfectly) the uncertainty associated with any individual RAP forecast.

Contrary to the deterministic case, where the values of the parameters involved in the phase-space reconstruction can be estimated *a priori* from properties of the time series of the observations^{25,26}, some parameters in RAP are treated as free, and are determined by maximizing or minimizing some objective function quantifying the discrepancy between forecasts and observations. For forecasts of the first kind, one would try to set the parameters in such a way as to minimize the difference between the forecasts and the observations. But if the forecasting systems, such as RAP, is a stochastic one, then distinct realizations of the same forecast may be quite different from each other, and the significance of an individual forecast's good or bad match of the observations can only be assessed in a statistical sense. Ideally, the observations and a *typical* member of a large ensemble of forecasts should be statistically indistinguishable. Our simple criterion of indistinguishability is based on the notion of *band depth*. In functional statistics, the band depth is a measure that generalizes order statistics to functional data, providing a way to rank curves based on their centrality or typicality within an ensemble of curves²⁹. It essentially quantifies how much of a curve's graph lies within bands formed by other curves in the sample. A higher band depth indicates a curve that is more central in the bundle of curves of the ensemble^{29,30}. According to this criterion, an ensemble of forecasts is considered to be a good match of the observations if the observations have a band depth close to the highest band depth of the individual members of the ensemble. When this happens, the graph of the observations remains central within the ensemble of forecasts.

The use of band depth also gives an objective way to single out the 'most typical' member of an ensemble: just as the median or the mean can be used as the individual number most representative of a set of numbers, likewise the forecast having the highest band depth in an ensemble can be taken as the individual representative of that ensemble²⁹, and used as if it were a deterministic forecast.

The rest of the paper is organized as follows: the 'Results' section compares hindcasts produced with the RAP algorithm with

SIE observations data reported by NSIDC. In the 'Discussion' section we summarize RAP's merits, highlight its weaknesses and make the case for using RAP as the benchmark of choice for quantifying the skill of SIE forecast systems. We refer the reader to the 'Methods' section for further details on the RAP algorithm and on how its parameters were selected.

Results

Typical results that can be obtained by RAP seasonal-scale forecasts of SIE are illustrated in Figure 1, showing 150-days long hindcasts encompassing the September minimum for 6 years spanning the satellite era. Each hindcast begins on June 1st. Here, and in any other hindcast in this paper, the NSIDC SIE data corresponding to the period being hindcasted is expunged from the dataset used in the embedding. Thus, as in an actual forecast, the set of analogs available to the algorithm does not contain what the algorithm is trying to predict.

Each panel in the figure shows the 10th to 90th percentile range of a set of 1000 RAP realization (gray bands); 10 individual RAP realizations, randomly chosen out of that set (thin gray lines); the RAP realization having the maximum band depth (thick blue line); and the lowess-smoothed NSIDC SIE data (thick orange line). The 10th to 90th percentile range generally has a width smaller than $2 \cdot 10^6 \text{ km}^2$ even on the 5th month of forecast, and it is often smaller than 10^6 km^2 . The individual RAP realizations show a realistic and believable SIE dynamics, with a rapid decline in July and August, a minimum in September and a sharp growth in October. In some years the highest band-depth RAP realization, is remarkably close to the NSIDC SIE data, and in general the latter fall within the 10th to 90th percentile range of the RAP forecasts. The notable exception is 2012, the sole year for which RAP's performance is poor. The exceptional sea ice loss observed in the Summer 2012 has been linked to a singular sequence of rare events (including the Great Arctic Cyclone of early August 2012) unlike anything that was observed since 1979^{31–33}. As such, the conditions in the Arctic during the summer of 2012 have no analogue in the record of the satellite era, and thus it must be expected that a forecast algorithm based on analogues would fail to reproduce that minimum.

The time series of the SIE September Average is shown in Figure 2. Note that RAP cannot compute an initial state corresponding to June 1st of the years 1979, 1980, 1988, 1989 because the delayed components of such state vector would either extend beyond the beginning of the dataset, or fall in the NSIDC SIE data gap covering the interval from 1987/12/02 to 1988/01/13 (see 'Methods' for additional details). Thus hindcasts for those years were not performed. The gray band covers the 10th to 90th percentile range of the September average SIE of a set of 1000 RAP realizations. The median of those realization is shown by the purple line. The September average of the realization with the highest band depth is shown by the blue line. The September average of the NSIDC SIE data is shown by the orange line. The observations fall within the 10th to 90th percentile range of the RAP hindcasts, with notable deviations in 1986, 1996, 1999, 2012. Except than in 1999 and 2012, the September average of the deepest RAP hindcast is always within 10^6 km^2 from the observed September average, and it is often much closer. The uncertainty quantified by the width of the 10th to 90th percentile range is generally smaller than $\pm 10^6 \text{ km}^2$. This is comparable with the performance of both dynamical forecasting systems³⁴ and of modern deep-learning forecasting systems^{35,36}. RAP's error in forecasting the September Average SIE is also much smaller than the spread among the models participating in the Sea Ice Outlook intercomparison of seasonal SIE forecasts organized by the Sea Ice Prediction Network^{37,38}. Finally, the RAP hindcasts nicely follow the overall trend of summer SIE decline, which progressed very rapidly in the early 2000's but has substantially decelerated in the last 15 years^{39,40}.

We use three standard indicators of skill⁴¹ to evaluate the performance of RAP in forecasting individual months with up to 9 months of lead time (Figure 3). Figure 3A shows the bias, or model error: the average difference between observations and forecast over the target month. Although negative and positive biases appear to cluster together, in no case the bias is significantly different from zero at 95% confidence level. Thus RAP forecasts are overall equally likely to overestimate or to underestimate the observations. Figure 3B shows the root mean squared error (RMSE). Combinations of lead time and target month leading to RMSE values significantly larger than $0.2 \cdot 10^6 \text{ km}^2$ at 95% confidence level are marked with a dot. The largest errors occur in predicting September and October with lead times larger than 3 months. In that case the RMSE can be as high as $0.6 \cdot 10^6 \text{ km}^2$. For forecast targeting winter and spring months the RMSE remains smaller than $0.4 \cdot 10^6 \text{ km}^2$, even with 9 months of lead time. Figure 3C shows the anomaly correlation coefficient (ACC). This indicator is the value of Pearson's correlation coefficient between the observed and the forecast anomaly with respect to the climatology. It measures the ability to correctly forecast fluctuations with monthly time scale diverging from the climatological average. RAP does not perform well respect to this indicator, showing some weak skill only in June, July and October. This result suggests that little or no information on future month-scale fluctuations in SIE is encoded in the historical time series of SIE, and thus RAP is unable to reproduce them correctly because it is a data-driven algorithm based only on historical SIE values. Indeed, month-scale fluctuations are generally believed to be the result of mesoscale weather disturbances, and of other transient phenomena^{42,43}. Meteorological or climate models are in a much better position for accurately reproducing those transient phenomena, and may thus achieve a better monthly ACC value even if their bias and RMSE values may be worse than RAP's⁴¹.

Discussion

In this paper we developed a Random Analogue Prediction (RAP) algorithm for the forecasting of Arctic Sea Ice Extent on seasonal time scales. RAP is a stochastic, data-driven algorithm that produces ensembles of forecasts. By applying the notion of band–depth (a recently developed statistical technique that extends the notion of centrality to functional data) we identify a representative realization in the ensemble that can be thought of as "best" forecast, and can be compared with deterministic forecasts. The rest of the ensemble is used to quantify the level of uncertainty associated with that "best" forecast.

We show that the skills of RAP may be competitive with those of state-of-the-art forecasting techniques, both based on physics (general circulation models) or on data-driven techniques much more sophisticated and data-hungry than RAP (deep-learning AI models). In particular, RAP has a negligible bias and a small RMS error ($0.6 \cdot 10^6 \text{ km}^2$ in the worst cases) for forecasts with up to 9 months of lead time. However, RAP lacks skill in faithfully reproducing short-term fluctuations in SIE (monthly scale or less) and we attribute this shortcoming to the data-driven nature of the algorithm: the historical SIE record does not contain the information necessary to forecast the meteorological phenomena that shape those short-term fluctuations.

In our opinion, RAP's biggest strength is its simplicity and interpretability. At its core, RAP is nothing else than the well known method of the analogs applied to the SIE time series, rather than to a spatially extended field. The prescription for defining the analogue states is inspired by results in the theory of dynamical systems that date back to the 1980s, and the selection of the specific analogue used to advance the forecast by one step follows a simple stochastic rule. Thus, RAP produces its forecasts by looking back in the SIE time series to find analogues of the present state, and extrapolates the observed evolution of the analogues to the current state.

We stress that the method of the analogues (and thus RAP) does not depend on underlying physical hypothesis, approximations or parameterizations. The only presupposition inherent in the method is the way in which the analog states are identified and selected. The few free parameters of the method are tuned in order to maximize the centrality (in the band–depth sense) of the observations with respect to an ensemble forecast covering the same time interval. Thus any skill exhibited by RAP only stems from its ability to properly apply to the present state changes that have been observed to occur in the past.

Simplicity, interpretability, independence from physical hypotheses, and the ability to produce ensembles that estimate the forecast error, in our opinion, make RAP the benchmark of choice that more complicated models should be compared against. In particular, a proper model (as opposed to a benchmark) should be able to consistently produce more information about the future than what is already encoded in the SIE data. In other words, those models should be able to draw additional information from the knowledge of e.g. atmospheric and ocean conditions, as well as from the laws of physics, to consistently beat a benchmark that just reproduces past dynamics observed in SIE data, and is oblivious of anything else.

Methods

We use a modified version of the Random Analogue Predictor (RAP) algorithm² applied to NSIDC sea ice extent data⁴⁴. The sea ice extent time series is interpolated and resampled with a constant sampling rate of f_s days with a lowess smoother using 2nd-order polynomials and window length of W days. Smoothing and resampling are intended to filter out the effect short-term weather events (e.g. synoptic-scale depressions) and yield a time series $s_1, \dots, s_n$. Gaps in the original time series (e.g. between December 2nd, 1987 and January 13th, 1988) are not filled by the lowess smoothing. The corresponding s_i values are marked as invalid.

The time series is then *embedded* in $\mathbb{R}^{M+1}$ (the embedding space), namely, it is transformed in the set of $(M+1)$ –dimensional state vectors defined as

$$\mathbf{x}_i = (s_i, s_{i-\tau}, \dots, s_{i-(M-2)\tau}, s_{i-(M-1)\tau} \sin(\varphi_i), s_{i-(M-1)\tau} \cos(\varphi_i)) \quad (1)$$

where the embedding delay τ is expressed as an integer multiple of the sampling rate, and the phase $\varphi_i = T^{-1}2\pi i f_s$ encodes the information that the time series, albeit non-periodic, is phase-locked to the year length $T = 365.25$ days. Phase-encoding is the only difference with the standard embedding procedure^{20,24}. Vectors containing one or more invalid components are considered as invalid, and are not used in the forecast.

A single random analogue forecast is performed with the following steps

1. select an initial state $\mathbf{y}_0$, which may or may not be in the set $\{\mathbf{x}_i\}$;
2. find the set $\mathcal{N}$ of all the valid state vectors, within a distance λ from $\mathbf{y}_0$ and having valid successors (the successor of $\mathbf{x}_i$ is $\mathbf{x}_{i+1}$); if there are no such state vectors, then $\mathcal{N}$ contains the closest valid vector to $\mathbf{y}_0$ having a valid successor.
3. randomly select the element $\mathbf{x}_{i_j}$ from $\mathcal{N}$, with a probability inversely proportional to the Euclidean distance from $\mathbf{y}_0$
4. define the forecast state $\mathbf{y}_1 = \mathbf{y}_0 + (\mathbf{x}_{i_j+1} - \mathbf{x}_{i_j})$

5. set $\mathbf{y}_1$ as the new initial state and repeat from 2. as many times as desired.

The method depends on the free parameters f_s , W , M , τ , λ . We set $f_s = 5$. This insures that the average distance between successive state vectors is substantially larger than the average distance between closest neighboring vectors. Insuring that states nearby in the embedding space are not also close in time has been shown to avoid biases in the geometry of the embedding⁴⁵. We set λ to be twice the median nearest neighbor distance of state vectors. This gives the RAP algorithm enough choice between alternative analogs, without allowing the choice of states that are too far from the current one, and thus may not be good analogs. In order to determine W, M, τ , we exhaustively explored the parameter space, setting: $W \in \{15, 30, 45, 60, 75, 90\}$ (days); $M \in \{2, 3, \dots, 12\}$; $\tau \in \{6, 12, 18, 24, 30, 36, 42, 48, 54, 60\}$ (resampling units). For each possible combination of the parameters, we ran 1000 RAP hindcasts starting on June 1st of the years 1979, ..., 2024 and extended for 150 days. To maintain the algorithm unaware of what actually happened in reality, we discarded the NSIDC data corresponding to the hindcasted time interval. To simulate the uncertainty due to short-term meteorological conditions and observational error, the initial state of each hindcast is perturbed with a vector whose components are random variables uniformly distributed in ± 0.2 millions of square Km. Each of the hindcast and the observations relative to a given year were then ranked (from lower to higher) according to their band-depth^{29,30} computed with the *statsmodel* v0.14.1 library⁴⁶. The ensemble of hindcasts for that year was then assigned a score equal to the band-depth rank of the observations divided by the number of hindcasts in the ensemble: a low score indicates that the observations are outliers in the ensemble of hindcasts, and a high score indicates that the observations fall in the center of the bunch of hindcasts. We selected the parameter set that yields the highest median score over all of the years covered by the NSIDC SIE dataset. This turns out to be $W = 15$ (days), $M = 5$, $\tau = 36$ (resampling units), having median score of 0.65, which means that for this parameter set, in half of the years the observations have a band depth higher than 65% of the hindcast curves. For comparison, the parameter set $W = 60$ (days), $M = 9$, $\tau = 60$ (resampling units) has the lowest median score of 0.17.

In order to evaluate the skills of the RAP forecast we use three standard indicators: mean error (Bias), root-mean-square error (RMSE), and anomaly correlation coefficient (ACC), defined as follows:

$$\text{Bias} = \frac{1}{n} \sum_{i=1}^n (y_i - s_i) \quad (2)$$

$$\text{RMSE} = \frac{1}{n} \sqrt{\sum_{i=1}^n (y_i - s_i)^2} \quad (3)$$

$$\text{ACC} = \frac{\sum_{i=1}^n (y_i - c_i)(s_i - c_i)}{\sqrt{\sum_{i=1}^n (y_i - c_i)^2 \sum_{i=1}^n (s_i - c_i)^2}} \quad (4)$$

where y_i , s_i , c_i , for $i = 1, \dots, n$, respectively are the forecasts, the observations, and the climatological averages for the period of interest.

References

1. Lorenz, E. N. Climatic predictability. In *The physical basis of climate and climate modelling*, vol. 16 of *GARP Publication Series*, 132–136 (WMO, 1975).
2. Paparella, F., Provenzale, A., Smith, L. A., Taricco, C. & Vio, R. Local random analogue prediction of nonlinear processes. *Phys. Lett. A* **235**, 233–240, DOI: [https://doi.org/10.1016/S0375-9601\(97\)00607-5](https://doi.org/10.1016/S0375-9601(97)00607-5) (1997).
3. Krishnamurthy, V. Predictability of weather and climate. *Earth Space Sci.* **6**, 1043–1056 (2019).
4. Boffetta, G., Crisanti, A., Paparella, F., Provenzale, A. & Vulpiani, A. Slow and fast dynamics in coupled systems: A time series analysis view. *Phys. D: Nonlinear Phenom.* **116**, 301–312 (1998).
5. Boffetta, G., Cencini, M., Falcioni, M. & Vulpiani, A. Predictability: a way to characterize complexity. *Phys. reports* **356**, 367–474 (2002).
6. Mitchell, L. & Gottwald, G. A. On finite-size lyapunov exponents in multiscale systems. *Chaos: An Interdiscip. J. Nonlinear Sci.* **22** (2012).
7. Ghil, M., Groth, A., Kondrashov, D. & Robertson, A. W. Extratropical sub-seasonal to seasonal oscillations and multiple regimes: the dynamical systems view. In *Sub-seasonal to seasonal prediction*, 119–142 (Elsevier, 2019).

8. Legras, B. & Ghil, M. Persistent anomalies, blocking and variations in atmospheric predictability. *J. Atmospheric Sci.* **42**, 433–471 (1985).
9. Mason, B. J. Numerical weather prediction. *Contemp. physics* **27**, 463–472 (1986).
10. Saha, S. & Van den Dool, H. A measure of the practical limit of predictability. *Mon. weather review* **116**, 2522–2526 (1988).
11. Van den Dool, H. M. *Empirical methods in short-term climate prediction* (Oxford University Press, 2007).
12. Palmer, T. & Tibaldi, S. On the prediction of forecast skill. *Mon. weather review* **116**, 2453–2480 (1988).
13. Toth, Z. Long-range weather forecasting using an analog approach. *J. climate* **2**, 594–607 (1989).
14. Lorenz, E. N. Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmospheric Sci.* **26**, 636–646 (1969).
15. Van den Dool, H. A new look at weather forecasting through analogues. *Mon. weather review* **117**, 2230–2247 (1989).
16. Takens, F. Detecting strange attractors in turbulence. In Rand, D. & Young, L.-S. (eds.) *Dynamical Systems and Turbulence, Warwick 1980: proceedings of a symposium held at the University of Warwick 1979/80*, vol. 898 of *Lecture Notes in Mathematics*, 366–381 (Springer, 1981).
17. Packard, N. H., Crutchfield, J. P., Farmer, J. D. & Shaw, R. S. Geometry from a time series. *Phys. review letters* **45**, 712 (1980).
18. Broomhead, D. S. & King, G. P. Extracting qualitative dynamics from experimental data. *Phys. D: Nonlinear Phenom.* **20**, 217–236 (1986).
19. Farmer, J. D. & Sidorowich, J. J. Predicting chaotic time series. *Phys. review letters* **59**, 845 (1987).
20. Sauer, T., Yorke, J. A. & Casdagli, M. Embedology. *J. statistical Phys.* **65**, 579–616 (1991).
21. Smith, L. A. Identification and prediction of low dimensional dynamics. *Phys. D: Nonlinear Phenom.* **58**, 50–76 (1992).
22. Smith, L. A. Local optimal prediction: exploiting strangeness and the variation of sensitivity to initial condition. *Philos. Transactions Royal Soc. London. Ser. A: Phys. Eng. Sci.* **348**, 371–381 (1994).
23. Stark, J., Broomhead, D. S., Davies, M. E. & Huke, J. Takens embedding theorems for forced and stochastic systems. *Nonlinear Analysis: Theory, Methods & Appl.* **30**, 5303–5314 (1997).
24. Stark, J. Observing complexity, seeing simplicity. *Philos. Transactions Royal Soc. London. Ser. A: Math. Phys. Eng. Sci.* **358**, 41–61 (2000).
25. Kantz, H. & Schreiber, T. *Nonlinear time series analysis* (Cambridge university press, 2003).
26. Bradley, E. & Kantz, H. Nonlinear time-series analysis revisited. *Chaos: An Interdiscip. J. Nonlinear Sci.* **25** (2015).
27. Aeyels, D. Generic observability of differentiable systems. *SIAM J. on Control. Optim.* **19**, 595–603 (1981).
28. Muldoon, M. R., Broomhead, D. S., Huke, J. P. & Hegger, R. Delay embedding in the presence of dynamical noise. *Dyn. Stab. Syst.* **13**, 175–186 (1998).
29. López-Pintado, S. & Romo, J. On the concept of depth for functional data. *J. Am. statistical Assoc.* **104**, 718–734 (2009).
30. Sun, Y., Genton, M. G. & Nychka, D. W. Exact fast computation of band depth for large functional datasets: How quickly can one million curves be ranked? *Stat* **1**, 68–74 (2012).
31. Zhang, J., Lindsay, R., Schweiger, A. & Steele, M. The impact of an intense summer cyclone on 2012 arctic sea ice retreat. *Geophys. Res. Lett.* **40**, 720–726 (2013).
32. Parkinson, C. L. & Comiso, J. C. On the 2012 record low arctic sea ice cover: Combined impact of preconditioning and an august storm. *Geophys. Res. Lett.* **40**, 1356–1361 (2013).
33. Jeong, H., Park, H.-S., Stuecker, M. F. & Yeh, S.-W. Record low arctic sea ice extent in 2012 linked to two-year la niña-driven sea surface temperature pattern. *Geophys. Res. Lett.* **49**, e2022GL098385 (2022).
34. Batté, L. *et al.* Summer predictions of arctic sea ice edge in multi-model seasonal re-forecasts. *Clim. Dyn.* **54**, 5013–5029 (2020).
35. Andersson, T. R. *et al.* Seasonal arctic sea ice forecasting with probabilistic deep learning. *Nat. communications* **12**, 5124 (2021).
36. Wang, W., Yang, W., Wang, L., Wang, G. & Lei, R. Seasonal forecasting of pan-arctic sea ice with state space model. *npj Clim. Atmospheric Sci.* **8**, 172 (2025).

37. Blanchard-Wrigglesworth, E. *et al.* Forecast skill of the arctic sea ice outlook 2008–2022. *Geophys. Res. Lett.* **50**, e2022GL102531 (2023).
38. Bushuk, M., Bhatt, U. S., Meier, W., Blanchard-Wrigglesworth, E. & Taylor, A. Sea ice outlook: 2025 june report. <https://www.arcus.org/sipn/sea-ice-outlook/2025/june> (2025). Accessed: 2025-09-22.
39. Francis, J. A. & Wu, B. Why has no new record-minimum arctic sea-ice extent occurred since september 2012? *Environ. research letters* **15**, 114034 (2020).
40. England, M., Polvani, L., Screen, J. & Chan, A. Minimal arctic sea ice loss in the last 20 years, consistent with internal climate variability. *Geophys. Res. Lett.* **52**, e2025GL116175 (2025).
41. Harnos, K., L'Heureux, M., Ding, Q. & Zhang, Q. Skill of seasonal arctic sea ice extent predictions using the north american multimodel ensemble. *J. Clim.* **32**, 623–638 (2019).
42. Agarwal, S., Moon, W. & Wettlaufer, J. S. Trends, noise and re-entrant long-term persistence in arctic sea ice. *Proc. Royal Soc. A: Math. Phys. Eng. Sci.* **468**, 2416–2432 (2012).
43. Finocchio, P. M. & Doyle, J. D. Summer cyclones and their association with short-term sea ice variability in the pacific sector of the arctic. *Front. Earth Sci.* **9**, 738497 (2021).
44. NSIDC. Sea ice index, version 4. <https://nsidc.org/data/g02135/versions/4>, DOI: [10.7265/a98x-0f50](https://doi.org/10.7265/a98x-0f50) (1978–2025). Accessed: 2025-09-30.
45. Theiler, J. Spurious dimension from correlation algorithms applied to limited time-series data. *Phys. review A* **34**, 2427 (1986).
46. Seabold, S. & Perktold, J. Statsmodels: econometric and statistical modeling with python. In Varoquaux, G., Millman, K. J. *et al.* (eds.) *Proceedings of the 9th Python in Science Conference*, 92–96 (Scientific Computing with Python Austin, TX, 2010).

Acknowledgements

F.P. acknowledges several useful discussions on this topic with prof. Marcello Vichi, University of Cape Town.

Funding

F.R. was supported by Tamkeen under the NYU Abu Dhabi Research Institute Award CG0002 to Stability, Instability, and Turbulence (SITE).

F.P. was supported by Tamkeen under the NYU Abu Dhabi Research Institute Award CG0009 to Mubadala Arabian Center for Climate and Environmental Sciences (ACCESS).

Author contributions statement

F.R. and F.P. contributed equally to this work.

Additional information

Accession codes The software used to produce the results discussed in this paper is publicly available on Zenodo: <https://doi.org/10.5281/zenodo.17348912>

Competing interests The authors declare no competing interests.

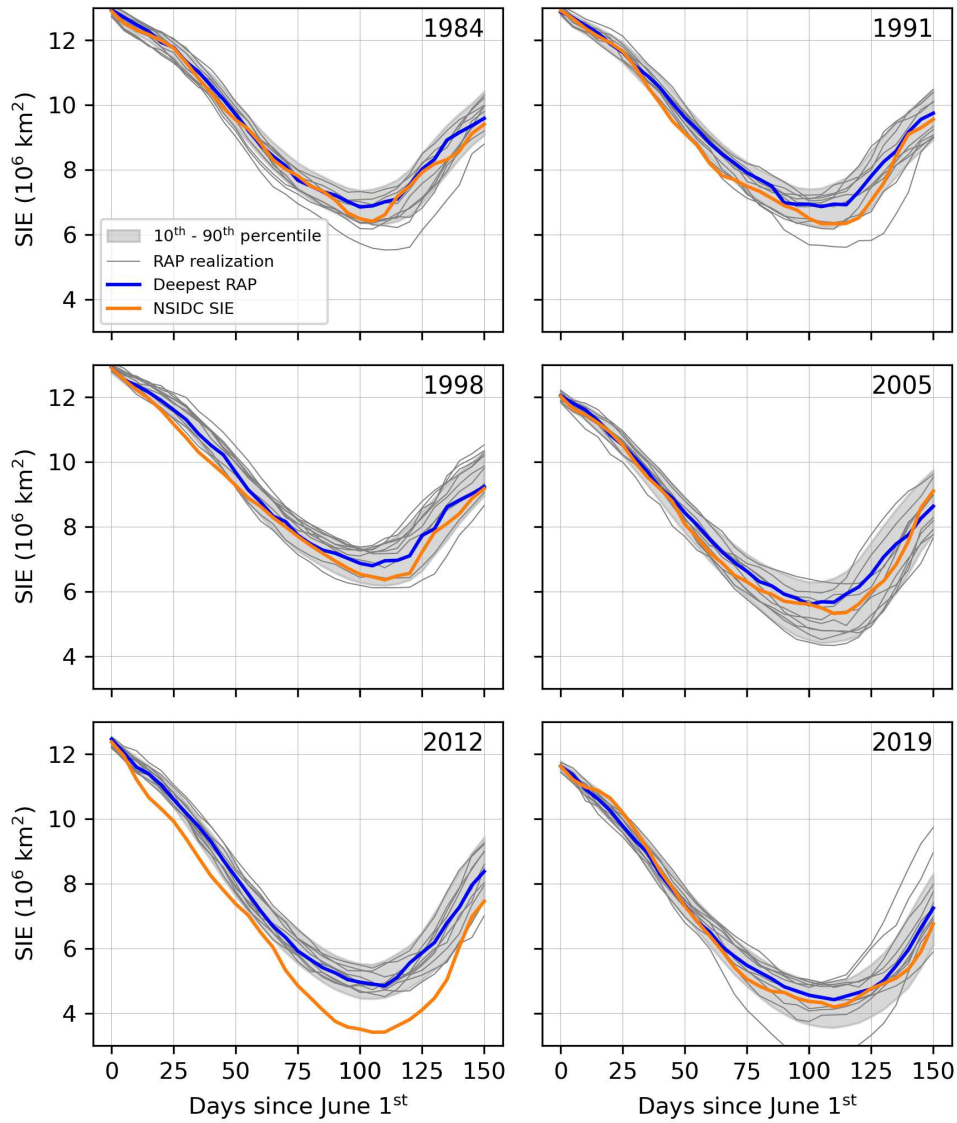


Figure 1. Hindcast RAP seasonal forecasts of Arctic sea-ice extent for the years 1984, 1991, ..., 2019. In all cases, the forecasts begin on June 1st and extend for 150 days. Orange line: NSIDC sea ice extent; these data are omitted from the dataset used to perform the RAP forecast relative to the same year. Blue line: RAP forecast with the highest band depth out of 1000 RAP realizations. Dark gray lines: 10 randomly chosen RAP forecasts. Gray band: 10th to 90th percentile estimated with 1000 RAP realizations

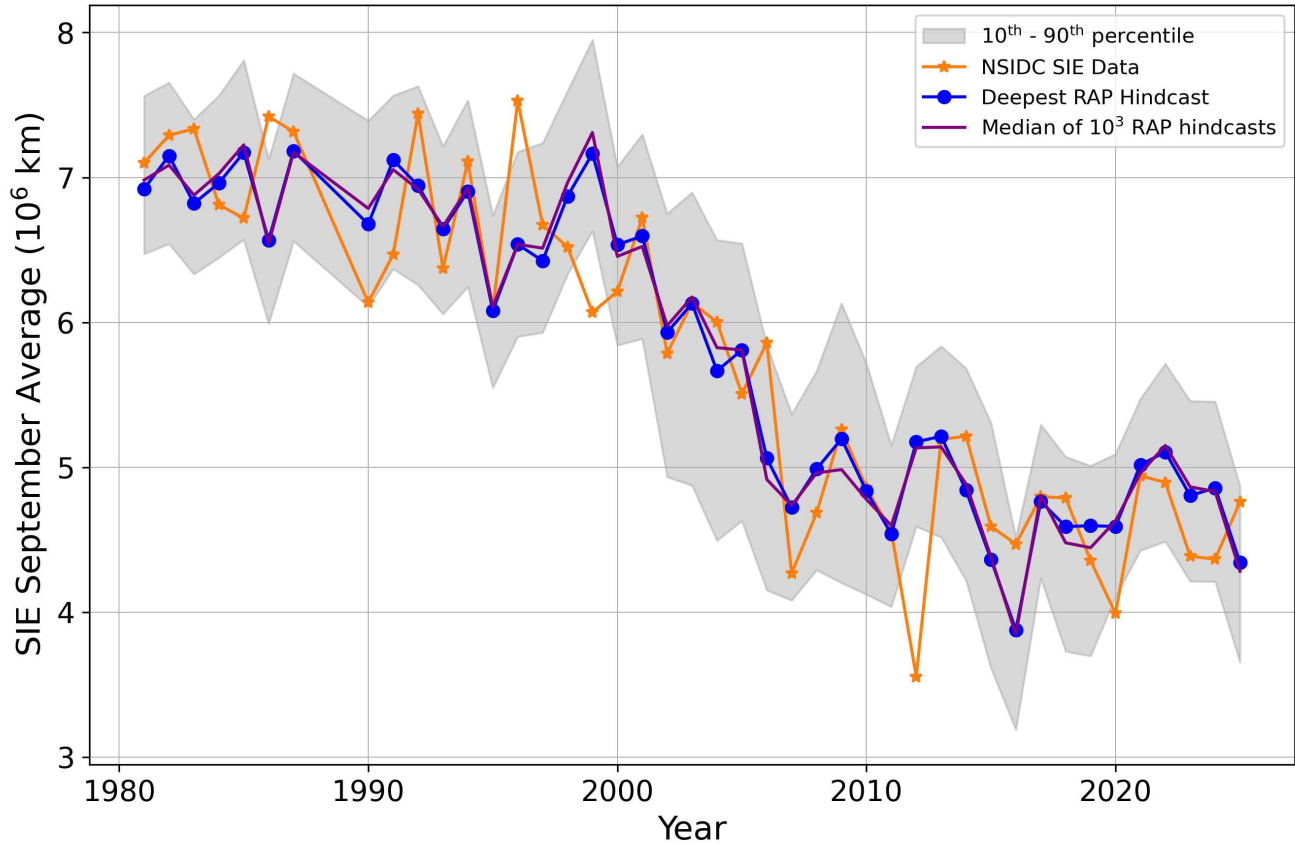


Figure 2. Time evolution of the September sea ice extent (SIE): NSIDC data (orange starred line) vs RAP hindcast (blue dotted line). The RAP September average is computed on the RAP forecast that has the highest band depth out of 1000 realizations. The purple line and the gray band mark, respectively, the median and the 10th to 90th percentile interval of the September average of those 1000 realizations. NSIDC SIE data for 2025 is not yet available at the moment of writing.

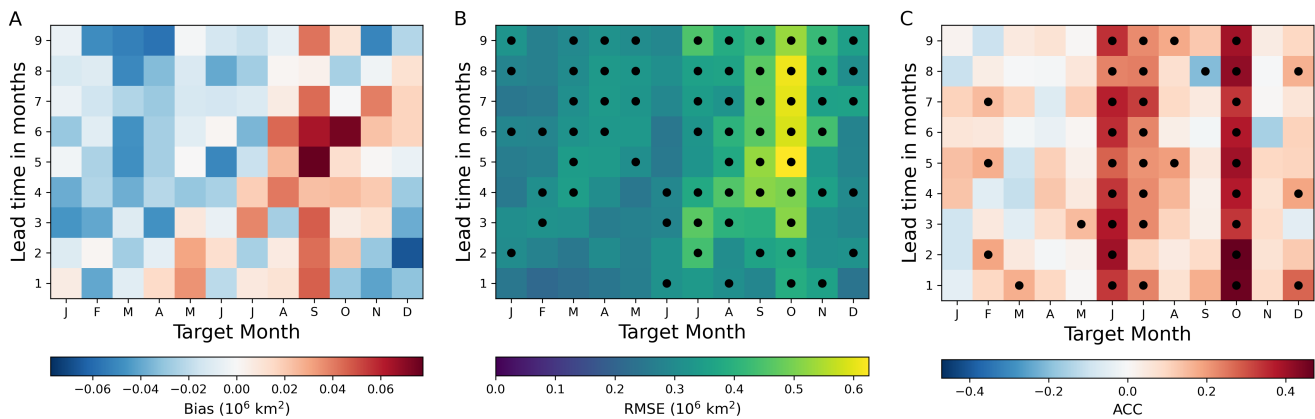


Figure 3. Indicators of forecast skill as a function of the forecast month and the lead time. Shown are the average values over all years in the NSIDC SIE dataset. Panel A: Bias (forecast minus observation); in no case the bias is significantly different from zero (95% t-test). Panel B: RMSE (root mean square error); the dots denote values significantly higher than $0.2 \cdot 10^6 \text{ km}^2$ (95% one-sided sign test). Panel C: ACC (anomaly correlation coefficient); the dots denote values significantly different from zero (95% t-test).