Robust Algorithms for Capturing Population Dynamics and Transport in Oceanic Variables along Drifter Trajectories using Linear Dynamical Systems with Latent Variables

Yan Yan^{2*¶}, Tony Jebara^{1,2*¶}, Ryan Abernathey^{3*¶}, Joaquim Goes^{3*¶}, Helga Gomes^{3*¶}

¹Department of Computer Sciences, Columbia University, New York, NY, USA

²Data Science Institute, Columbia University, New York, NY, USA

³Lamont Doherty Earth Observatory, Columbia University, Palisades, New York, NY, USA

* Corresponding authors:

E-mail: yy2250@columbia.edu (YY), jebara@cs.columbia.edu (TJ), rpa@ldeo.columbia.edu (RA),
jig@ldeo.columbia.edu (JG), helga@ldeo.columbia.edu (HG)

¶ These authors contributed equally to this work.

Abstract

The blooms of *Noctiluca* in the Gulf of Oman and the Arabian Sea have been intensifying in recent years posing a threat to regional fisheries and the long-term health of an ecosystem supporting a coastal population of nearly 120 million people. We present the results of a microscopic data analysis to investigate the onset and patterns of the *Noctiluca* (mixotrophic dinoflagellate *Noctiluca scintillans*) blooms, which form annually during the winter monsoon in the Gulf of Oman and in the Arabian Sea. Our approach combines methods in physical and biological oceanography with machine learning techniques. In particular, we present a robust algorithm, the variable-length Linear Dynamic Systems (**vLDS**) model, that extracts the causal factors and latent dynamics at the microscopic population-level along each individual drifter trajectory, and demonstrate its effectiveness by using it to test and confirm previously benchmarked macroscopic scientific hypotheses. The test results provide microscopic statistical evidence to support and recheck the macroscopic physical and biological Oceanography hypotheses on the *Noctiluca* blooms; it also helps identify complementary microscopic dynamics that might not be visible or discoverable at the macroscopic scale. The vLDS model also exhibits a generalization capability (inherited from a machine learning methodology) to investigate important causal factors and hidden dynamics associated with ocean biogeochemical processes and phenomena at the population-level.

KEY WORDS: *Noctiluca* blooms; GDP Lagrangian drifters; Microscopic hypothesis testing; Variable-length Linear Dynamical Systems (vLDS); Machine learning and artificial intelligence; Latent variables; Expectation Maximization algorithm

Introduction

Background

Recent advances in Data Science and Machine Learning have produced great successes in a variety of data-driven modeling for interdisciplinary scientific problems concerning complex natural phenomena, in Marine Ecology [1-6], Climatology [7], Oceanography [8-11], Geoscience [12], Computer Vision [13-15], Social Science [16], Computational Neuroscience [17-20], Speech and Language Processing [21-23], and Environmental Health Science [24]. Here we use these techniques to investigate the onset and patterns of the *Noctiluca* winter monsoon blooms, which form annually and with predictable regularity in the Gulf of Oman and in the Arabian Sea. Our approach relies on a combination of satellite and drifter derived oceanographic data and machine learning techniques. In particular, we obtain a robust model (the vLDS mode, variable-length Linear Dynamic System Model) that is capable of identifying the causal factors and dynamics at the microscopic population-level along each individual drifter trajectory. Furthermore, we assess its effectiveness via testing and confirm previously benchmarked scientific hypotheses. Rigorously statistical, the vLDS model is a powerful tool that helps us: 1) discover microscopic causal relationships in a high-dimensional dataset and 2) identify complementary microscopic dynamics that might not be discoverable at the macroscopic scale or accessible in controlled laboratory experiments.

The significance of this research is that these blooms of *Noctiluca* have been intensifying in recent years posing a threat to regional fisheries and the long-term health of an ecosystem supporting a coastal population of nearly 120 million people [25-28]. When seen from space, the *Noctiluca* blooms appear as large drifting swirls and filaments on the surface of the sea (Fig 1A). Traditionally, photosynthetic diatoms supported the Arabian Sea food chain; zooplankton preyed primarily on diatoms, which were in turn grazed by fish. Since early 2000s, the ecosystem of the Arabian Sea

appears to have changed [27], with large and widespread blooms of *Noctiluca* superseding diatoms which dominated winter monsoon blooms. Gomes et al. [26] were able to discover that the annual outbreaks of *Noctiluca* were linked to the up shoaling of hypoxic waters from depth into the euphotic zone. Within a decade and half, *Noctiluca* blooms have virtually replaced diatoms at the base of the food chain, marking what appears to be an unprecedented ecosystem shift.

In previous studies, Gomes et al. [25-26], Goes et al. [28] relied on satellite observations and in-situ data sampling and biologically controlled experiments on board of research vessels, to describe the underlying dynamics governing the transport, the growth and decay of the *Noctiluca* blooms in the Arabian Sea region. It has been demonstrated that *Noctiluca* is capable of migrating up and down in the water column depending on conditions at the surface. Like all dinoflagellates [29-32], a stable water column is essential for the growth and proliferation of *Noctiluca* as large surface blooms. Since it is a mixotroph, it can meet its metabolic requirements via feeding on external prey or through photosynthesis by thousands of green “endosymbionts,” living within its central symbiosome (Fig 1B). This flexibility gives it an edge on diatoms, which survive on sunlight alone. When feeding on an external source of organic matter *Noctiluca* tends to accumulate significant amounts of ammonia [33] and lipids [27], making them highly buoyant, allowing them to accumulate at the sea surface for prolonged periods of time. When present at the surface *Noctiluca* can be easily transported by physical oceanographic processes, where there appear in satellite ocean color in association with micro and meso-scale eddies, filaments and streamers. Over the past decade in particular, *Noctiluca* blooms have become more intense and widespread. The blooms of 2015 and 2017 were the largest on record occupying an area almost thrice the size of the State of Texas.

Summarizing this complex biogeochemical process, recent research [26-27] provided indications that the biological or environmental trigger for the *Noctiluca* blooms were nutrient rich, low-oxygen waters, brought to the surface by an annually recurring mesoscale cyclonic eddies. The eastward propagation of these eddies and filaments associated with them appear to be responsible for the gradual spread of the bloom from their inception in the Sea of Oman in Dec. towards the east by January until mid-March (Fig 1C). The underlying forces responsible for the dispersal of the winter-time *Noctiluca* blooms are multiple. While *Noctiluca's* unique physiological properties play an important role in maintaining the blooms at the surface and the overall integrity of the bloom, its dispersal is primarily due to ocean currents and the directions in which they move and mix with other water masses [34-39]. This mixing at meso- and sub-mesoscale occurs and changes over a seasonal basis, due to the interactions between the ocean and atmosphere that are driven by the monsoonal wind forcing [40-41].

Fig 1. (A) Satellites images displaying *Noctiluca* blooms as large swirls on the surface of the Arabian Sea [42]. **(B)** *Noctiluca scintillans* with a flick of its tail-like flagellum drawing energy from the millions of green algae, or “endosymbionts”, captured inside its transparent cell walls. **(C)** Satellite image for the monthly data of the chlorophyll *a* concentration in the Arabian Sea during year 2013, 2015, and 2016.

Based on our previous research [26] and given *Noctiluca's* rise in the Arabian Sea to an oxygen-starved hypoxia zone of more than 200 thousand square miles, we first (1) hypothesized and tested that nutrient-enrichment of the surface waters are increasing productivity and expanding the hypoxia zone as oxygen gets depleted. The potential sources of nutrients are multiple [25-26]. They

include the domestic and industrial fallout from countries bordering the Arabian Sea, winter convective mixing during Northeast Monsoon (NEM, Nov–Jan) which helps bring up nutrients from the mixed layer into the euphotic layer of the Arabian Sea, and the atmospheric deposition measured by satellite as the aerosol optical thickness at wavelength (865 nm). As *Noctiluca* grows, it uses up oxygen in the water when feeding on other phytoplankton. Grazing on other phytoplankton and declining oxygen levels allows *Noctiluca* to outcompete diatoms during winter. We then (2) hypothesized and tested that *Noctiluca* grew faster in light than in dark on the sea surface and in the sea water, thanks to its sun-loving endosymbiotic algae, which are thought to have survived 1.3 billion years since being on an oxygen-scarce Earth.

Goal and outline

To understand the microscopic impact of the biological and physical oceanographic factors on *Noctiluca* blooms along drifter trajectories, we have collected, combined, and preprocessed data from both the Ocean and Satellite datasets. The trajectory of each drifter is recovered by its spatio-temporal information. The biological and physical oceanographic profiles associated with the spatio-temporal coordinates of each drifter is then utilized to discern the the behavior and movement of the *Noctiluca* blooms along the trajectory of the particular drifter. This behavior and movement is statistically learned or transformed by the vLDS model. Since the drifters have different launch time and longevity, the time series representing individual drifter trajectories have different lengths. It is our goal in this paper to present the variable-length Linear Dynamic Systems (vLDS) model that is tailored to this particular data structure, to learn, summarize, and recover the latent dynamics for all drifter trajectories, and to generate predictive dynamics that match most closely the observed data along drifter trajectories.

The effectiveness of the vLDS model is demonstrated by testing our previous two hypotheses in the “Background” Section from the microscopic level and on both the physical and biological dynamics of the *Noctiluca* blooms. Moreover, *a priori* knowledge of the unique trait of *Noctiluca* is used in this study. The satellite based datasets that most closely align with the growth of *Noctiluca* blooms are colored dissolved organic material (*CDOM*) content of sea water, photosynthetically available radiation and diffuse attenuation coefficient at 490 nm using the Lee algorithm (light on and under the sea surface) and aerosol optical thickness over water (dust over the sea), and 4-micron nighttime sea surface temperature (*SST4*). The physical dispersion of the drifters or particles on the ocean surface is described by its spatio-temporal coordinates, eastward and northward velocity components, speed, and distance to the coast. The analysis of the phytoplankton blooms in [25-26, 40-41] is on the macroscopic scale of space and time, namely, the data is aggregated or pooled across spatio-temporal dimensions. In this study, the surface velocity drifter dataset [43] and satellite image dataset [44-47] are not aggregated or pooled across spatio-temporal dimensions. With the data structure of individual drifter trajectories kept intact, the vLDS model tests and works under our previous hypotheses on the dynamics of the *Noctiluca* blooms from the microscopic scale of space and time in the Arabian Sea region. More specifically, the vLDS model summarizes and captures the statistical structure of all drifter trajectories, by maximizing the log-likelihood of the joint distribution of the observations and their latent state variables and allowing each individual drifter trajectory generated by the model to vary as smooth functions in the latent dynamical state. By comparing the vLDS model predictions directly with the observed ocean profiles along the drifter trajectories, we can easily visualize its predictive performance and interpret the underlying dynamics of the *Noctiluca* blooms at the microscopic scale, as discussed in the “Discussion & conclusion” Section.

The benefit of vLDS is threefold. First, it provides statistical evidence in a direct and zoomed-in manner to support and recheck the macroscopic physical and biological oceanographic observations and the inherent physiological behavior of *Noctiluca* blooms, by recovering the latent dynamics that governs the probabilistic distribution of the *Noctiluca* concentration in space and time and by comparing the model predictions with the observed ocean profiles. Our study shows that most of the macroscopic hypotheses on the *Noctiluca* blooms can be statistically validated by the vLDS model. Second and more importantly, it helps us to identify complementary microscopic dynamics that might not be visible or discoverable at the macroscopic scale. The vLDS predictive plots in “Discussion & Conclusion” Section provide statistical evidence that the atmospheric deposition measured by the quantities *T865* aerosol optical thickness at wavelength 865 nm does not have much impact on the underlying dynamics that are driving the *Noctiluca* growth that is measured by the chlorophyll *a* concentration (*Chl a*) at the time scale of two days. Third, it provides a generalizable machine learning methodology to probe important microscopic causal factors and hidden dynamics for the ocean biogeochemical processes at the population-level along individual drifter trajectories. These scientific findings in the microscopic scale of the population can lead to critical hypothesis and even conclusions at the macroscopic scale of the pooled data. We note that this microscopic approach has been successfully applied in other areas, such as computational neuroscience [19-20] and sound tracking [22].

The major assumption of the **vLDS** model is that all the drifter trajectories are governed by the same set of biological and physical oceanographic principles or dynamics, in other words, all drifter trajectories in the vLDS model share the same model parameters. This assumption is both natural and physical. Those model parameters can then be learned jointly from the data, even if each drifter trajectory is evolving independently under the shared parameters to capture the autocorrelation

and the latent dynamics in the full spatio-temporal, biological, and physical feature space. The model's ability to capturing the underlying physical and biological dynamics of the *Noctiluca* blooms is quantified by the R-squared (R^2) of the vLDS model fitted on the cross-validation dataset and predicted on both the cross-validation and heldout data. After the vLDS model parameters are learned from the data, we observed that the the vLDS model, recovers all the spatio-temporal variables well along drifter trajectories, as well as most of the biological factors, namely, the *CDOM*, light on and under the sea surface, *SST4*, except for the atmospheric deposition *T865*.

These test results support the claim that the assumption on the vLDS model is natural. Moreover, the recovered, summarized, and predicted latent dynamics (rather than the observed time series) has shown a close relationship among *Noctiluca* blooms, physical dispersal, and biological environments. The highly correlated relationships between the *Chl a* and *CDOM*, and between the *Chl a* and diffuse attenuation coefficient at 490 nm using the Lee algorithm (*KD490*, light under the sea surface) are close to linear. The tightly correlated relationships between the *Chl a* and photosynthetically available radiation (*PAR*, light on the sea surface) and between the *Chl a* and *SST4* are nonlinear. The aerosol optical thickness at wavelength 865 nm (*T865*) does not have much impact on the vLDS recovered dynamics or the *Chl a*. This corresponds to our hypotheses in the "Background" Section, and demonstrates that the nutrient and light are two important positive factors for the *Noctiluca* blooms and that the atmospheric deposition does not contribute much to the *Noctiluca*'s growth. Our two hypothesis are thus tested and confirmed by these test results.

Materials and methods

Data collection

The Arabian Sea (coordinate range ~ 5 to 28° N, 45 to 75° E) is predominantly located in the tropics (Fig 1A), and it has one of the most energetic current systems driven by the seasonally reversing monsoons. Due to the temperature gradient across the land mass and ocean, the winter monsoon from November to February is cold and dry and is blowing from the Indian subcontinent. Moreover, the winter monsoon cools the surface of the ocean and induce convective mixing that brings nutrients from depth to the surface. Nutrients and low oxygen waters brought to the surface by winter convective mixing during the winter monsoon foster large *Noctiluca* blooms of the Arabian Sea [26].

Datasets on *Chl a*, from the ocean color satellites Sea-WiFS, MODIS-Aqua (NASA), VIIRS (NOAA) and MERIS and OLCI-A (ESA), were used for studying the distribution of *Noctiluca* blooms during winter. In practice, the Ocean color datasets have missing values at certain locations due to the limitation of the satellite coverage (and also to the presence of clouds). The dark spots are regions with missing values (Fig 2). For the purpose of our study, we used merged products from both NASA and the GlobColour Project [44-47]. Ocean color satellites can provide remote sensing reflectance values for different wavebands. These wavebands are used in empirical and semi-analytical algorithms to convert remote sensing reflectance to chlorophyll *a* concentration. Pre-processed *Chl a* data products were used to explore a time series of snapshots of chlorophyll *a* concentration on a lattice of latitude and longitude coordinates. In the next sections, we provide detailed descriptions of the data aggregation and preprocessing steps.

Fig 2. Level-3 data for the chlorophyll *a* concentration from Dec. 27 to Dec. 31, 2015 [42]. The dark regions indicate missing values.

234

235 The temporal evolution of the satellite images reflects both physical and biological dynamics. To
 236 impose the structures of physical drivers (advection) onto the data sample, we utilized the drifter
 237 array data from the NOAA's Global Drifter Program (GDP). These freely drifting buoys provide
 238 information about the upper ocean currents that are responsible for the advection of the planktonic
 239 particles [48]. The Lagrangian trajectory of each float can be retrieved from the database as a time
 240 series of variables representing the location, velocity field, and sea surface temperature. We note
 241 that each float has a typical lifetime of a couple years and has different launch times. Therefore, the
 242 drifter dataset is highly heterogeneous in both time and space. In Fig 3, we display the temperature
 243 measurements of all the drifters in the Arabian Sea. It is known that the Lagrangian drifter
 244 trajectories are highly chaotic [49-51], and the prediction of particle trajectories has been a
 245 challenging research task [52]. There have been recent research results on using Kalman filter and
 246 data assimilation with various specific physical model, namely, the Gauss–Markov Lagrangian
 247 particle model [8], the Eulerian velocity field [9, 11], and the upper ocean horizontal momentum
 248 balance model from Ekman dynamics [10], to track the position and velocity of the floats. For
 249 comparison in our study, we are introducing and imposing statistical structure on the latent state
 250 variables to capture the joint dynamics among the *Chl a*, the spatio-temporal information of the
 251 floats, namely, the latitude, longitude, velocity, speed and distance to the coast, and the biological
 252 predictors, such as *CDOM*, *KD490*, *T865*, *PAR*, and *SST4*. The predictive plots generated by the vLDS
 253 model in our study (as displayed in Section “Discussion & Conclusion”) reveal the relationships
 254 among the physical, biological ocean profiles and the *Noctiluca* blooms inside a collection of chaotic
 255 drifter trajectories in the Arabian Sea region from 2002 to 2017.

256

257

258 **Fig 3.** Temperature trajectories of all drifters in the Arabian Sea Region from 2002 to 2017.

259

260 Combining multiple dataset

261 We merged the satellite data with the buoy data to generate a Lagrangian dataset. It is a collection
 262 of multivariate time series for each drifter with a unique id to combine the information from the
 263 satellites, namely, *Chl a*, *CDOM*, *KD490*, *T865*, *PAR*, and *SST4*, and the data associated with the drifters
 264 including *id*, *time*, latitude, longitude, velocity components, speed, and distance to the coast.
 265 Fourteen features were selected in our experiments (Table 1). The drifter *id* and *time* are mainly
 266 used for ordering and grouping data in the vLDS model. The twelve (12) remaining factors represent
 267 the physical and biological variables that related to the evolution of *Noctiluca* blooms [25-26]. Since
 268 the timescale of the phytoplankton reproduction is of the order of a few days [53], we carried out a
 269 resampling process to match the frequency of both datasets to the same level. At the same
 270 frequency, we interpolated the variables from the satellite dataset onto the specific spatial and
 271 temporal points of the Lagrangian drifter dataset, to add more information for each observation in
 272 the drifter dataset. This process builds up a multivariate time series for each drifter id with all the
 273 biological and physical information embedded on the drifter trajectory. Further, this interpolation
 274 process is repeated for all other features from the satellites.

275

276 **Table 1. Factors used in the vLDS model.**

Factor	Description	Median (range)
<i>id</i>	Drifter id	Totally 230 drifters with 5594 data records
<i>time</i>	Time of the observation	Nov 1 to Mar 31, from 2002 to 2017
<i>lat</i>	Latitude	12.92 °N (5.03 - 26.99)
<i>lon</i>	Longitude	63.86 °E (45.07 - 74.95)
<i>ve</i>	Eastward velocity component	-4.54 cm/s (-122.87 - 108.47)
<i>vn</i>	Northward velocity component	1.58 cm/s (-132.75 - 114.55)
<i>spd</i>	Speed of the drifter	22.14 cm/s (0.86 - 146.58)

<i>dist</i>	Distance to the nearest coast	390.85 km (1.33 – 1166)
<i>chlor_a</i>	Chlorophyll <i>a</i> concentration (<i>Chl a</i>)	0.25 mg m ⁻³ (0.03 – 44.77)
<i>sst4</i>	Nighttime sea surface temperature at 4-micron (<i>SST4</i>)	27.06 °C (0 – 30.06)
<i>cdm</i>	Colored dissolved and detrital organic materials (<i>CDOM</i>) absorption coefficient at 443 nm	0.02 m ⁻¹ (0.01 – 0.81)
<i>kd490</i>	Diffuse attenuation coefficient at 490 nm using the Lee algorithm (<i>KD490</i>)	0.07 m ⁻¹ (0.04 – 1.36)
<i>t865</i>	Aerosol optical thickness over water (T865)	0.13 (0.02 – 0.56)
<i>par</i>	Photosynthetically available radiation (<i>PAR</i>)	46.18 Einstein m ⁻² day ⁻¹ (12.20 – 57.80)

We note that each float record has information on its coordinates $\{lat, lon, time\}$. For convenience, we denote $\{lat, lon, time\}$ as $\{x, y, t\}$, which is almost always not precisely on the grid of the Ocean Color dataset. To resolve this issue, we now describe the multidimensional interpolation operator to map the chlorophyll *a* concentration onto the GDP float dataset. For any float data point with coordinates $\{x_0, y_0, t_0\}$, we identify the coordinate cube or the grid cell in the Ocean Color dataset that contains this point. Particularly, this cube has 8 vertices with coordinates generated by the outer product $\{x_{nearest}, x_{next}\} \otimes \{y_{nearest}, y_{next}\} \otimes \{t_{nearest}, t_{next}\}$, where the subscripts *nearest* and *next* indicate the nearest and the furthest neighbours in the cube for each coordinate. We further denote the interpolation weight w_x for x_0 as

$$w_x = \frac{x_0 - x_{nearest}}{x_{next} - x_{nearest}}$$

Similarly, we define the weights w_y and w_t for y and t . Using the function $f(x, y, t)$ to represent the chlorophyll *a* concentration at the coordinate $\{x, y, t\}$, we write the interpolated chlorophyll *a* concentration at $\{x_0, y_0, t_0\}$ as

$$\begin{aligned} f(x_0, y_0, t_0) = & (1 - w_x)(1 - w_y)(1 - w_t) \cdot f(x_{nearest}, y_{nearest}, t_{nearest}) \\ & + (1 - w_x)(1 - w_y)w_t \cdot f(x_{nearest}, y_{nearest}, t_{next}) \\ & + (1 - w_x)w_y(1 - w_t) \cdot f(x_{nearest}, y_{next}, t_{nearest}) \\ & + (1 - w_x)w_yw_t \cdot f(x_{nearest}, y_{next}, t_{next}) \end{aligned}$$

$$\begin{aligned}
 &+ w_x(1 - w_y)(1 - w_t) \cdot f(x_{next}, y_{nearest}, t_{nearest}) \\
 &+ w_x(1 - w_y)w_t \cdot f(x_{next}, y_{nearest}, t_{next}) \\
 &+ w_xw_y(1 - w_t) \cdot f(x_{next}, y_{next}, t_{nearest}) \\
 &+ w_xw_yw_t \cdot f(x_{next}, y_{next}, t_{next})
 \end{aligned}$$

In our implementation, we have applied the interpolation process to the chlorophyll *a* concentration, distance to the nearest coast, and all other predictors. As an illustration of the interpolation process, we display the distance to the nearest coast [42] with a resolution of 4km in Fig 4A and the interpolated distance to the nearest coast for all the floats in Fig 4B.

Fig 4. (A) Distance to the nearest coast for all the geographical locations in the Arabian Sea Region. **(B)** Interpolated values of the distance to the nearest coast for all the data points in the drifter dataset.

Using the multidimensional interpolation procedure described above, we map the satellite observations for each of the variables $\{ 'chlor_a', 'dist', 'cdm', 'kd490', 't865', 'par', 'sst4' \}$ onto the GDP floats. Along with the information on the floats, namely $\{ 'time', 'id', 'lat', 'lon', 've', 'vn', 'spd' \}$, the interpolated float dataset becomes high dimensional (Table 1). We remark that *'chlor_a'* denotes the chlorophyll *a* concentration, *'dist'* the distance from nearest coast, *'cdm'* the colored dissolved and detrital organic materials (*CDOM*) absorption coefficient at 443 nm, *'kd490'* the diffuse attenuation coefficient at 490 nm using the Lee's algorithm (*KD490*), *'t865'* the aerosol optical thickness over water (*T865*), *'par'* the photosynthetically available radiation (*PAR*), *'sst4'* the 4-micron nighttime sea surface temperature (*SST4*), *'id'* the id of a float, *'ve'* the eastward velocity component, *'vn'* the northward velocity component, and *'spd'* the speed of a float. These biological and physical factors are chosen to represent all the possible causes for the distribution of the *Noctiluca* blooms in the Arabian Sea. The *'chlor_a'* measured during the period from November 1 to March 31 are mostly

attributed to *Noctiluca* blooms. The ‘*cdm*’ measures *CDOM* the amount of dissolved organic materials in the sea water, which supports the growth of the *Noctiluca* [26]. The ‘*dist*’ measures the distance from the particle to the nearest coast, which is the source of the nutrient rich water. Moreover, during the winter season, the northwestern Arabian Sea off the coast of Oman experiences winter convective mixing (Nov–Jan), during which time nutrient rich, low-oxygen, cold water is brought to the surface both by convective mixing and by cyclonic eddy activity and benefit the growth of the *Noctiluca* in a complex and nonlinear fashion as described in the “Discussion & Conclusion” Section. The factor ‘*par*’ measures *PAR* the amount of light that is available on the sea surface for the photosynthesis of the symbiotic green algae (Fig 1B) living within *Noctiluca*. The ‘*kd490*’ provides an indication of the transparency of the water column and amount of light that penetrates into the sea water. The ‘*t865*’ represents the amount of particles in the atmosphere over the water, and is an indicator for the atmospheric deposition. The rest of the factors {‘*time*’, ‘*id*’, ‘*lat*’, ‘*lon*’, ‘*ve*’, ‘*vn*’, ‘*spd*’} are the spatio-temporal information describing the physical transport and dispersal of the *Noctiluca* blooms.

Data preprocessing for vLDS

Due to the limitation of the satellite coverage mentioned in the “Data Collection” Section, there are missing values in each of the variables in the interpolated dataset. Our focus here is on the chlorophyll *a* concentration ‘*chlor_a*’, since it is the key variable for *Noctiluca* blooms. The data structure of the post-processed float dataset is determined by the pre-processing steps for the variable ‘*chlor_a*’.

The overall goal of data preprocessing is to keep the microscopic trajectory-based data structure intact, split the drifter trajectories that are spanning over multiple years, and remove the drifter

trajectories that are too short for the learning algorithm, with the logical consideration of the physical and biological meaning. We consider that each period from November 1 to March 31 (during winter monsoon) represents one growth cycle of the *Noctiluca* bloom for a particular float or drifter. For any float with a unique id that has chlorophyll *a* data over two or more cycles, we split the data and assign a new derived float id to the data within each cycle, by adding a small increment 0.05 to the original float id. The resulting dataset allows the vLDS model to learn the microscopic dynamics of each cycle independently.

For each float id, we calculate the percentage of 'NaN' values in the dataset in the column '*chlor_a*' and choose a threshold of 40%. There are two cases. First, if this percentage is smaller than the threshold, the data quality for this particular float is considered to be good, and we interpolate all the missing values for each of the variables in {'*chlor_a*', '*dist*', '*cdm*', '*kd490*', '*t865*', '*par*', '*sst4*'}. See Fig 5 for an example, in which the time series is interpolated on '*chlor_a*'. Also, after the interpolation process, there might still be gaps in the time series. For instance, the float might just do not have any record, including 'NaN', in a certain short period. In this case, the float will be further split into continuous subseries. Therefore, every interpolated float time series will go through the second step, which we now describe, for checking and splitting.

Fig 5. The time series of '*chlor_a*' from the float id 64113560 on the left panel is interpolated and shown on the right panel.

In the other case, if this percentage of 'NaN' values is larger than the threshold, the data quality for this particular float is considered to unsuitable for interpolation. We split the discontinuous series into smaller continuous series on '*chlor_a*'. We loop through the time series on '*chlor_a*' and

split it into smaller continuous series on '*chlor_a*'. Moreover, we assign a new derived float id to each newly generated shorter series, by adding a small increment 0.03 to the original float id. Also, we drop any series on '*chlor_a*' of length 1. See Fig 6 for an example, in which we split the time series on '*chlor_a*' into 5 different shorter series, and we drop three series of length 1.

Fig 6. The time series of '*chlor_a*' from the float 71140 on the top panel is split into 5 different shorter series. Three series of length 1 are dropped, and the remaining two are shown on the bottom panel

Linear Dynamical Systems (LDS)

After being preprocessed, the merged GDP floats dataset consisted of 186 float records. For each float, the measurement is a multivariate time series $\mathbf{y}$, given by

$$\mathbf{y} := \{lat, lon, ve, vn, spd, dist, chlor_a, par, cdm, t865, kd490, sst4\}.$$

We use $\mathbf{x}^n = (\mathbf{x}_1^n, \mathbf{x}_2^n, \dots, \mathbf{x}_{T_n}^n)$ to denote the latent variables and $\mathbf{y}^n = (\mathbf{y}_1^n, \mathbf{y}_2^n, \dots, \mathbf{y}_{T_n}^n)$ the observations from the float n , and T_n for the length of the time series of the float n . The plain version Linear Dynamical System (LDS) is an adaptive procedure that can learn from data to recover the latent relationship between $\mathbf{y}$ and $\mathbf{x}$, using assumptions of linear relationships at time i between $\mathbf{x}_i$ and $\mathbf{x}_{i-1}$, $\mathbf{y}_i$ and $\mathbf{x}_i$, with Gaussian Noise. For the moment, we omit the superscript n and focus on one float. More specifically, we assume that for a specific float the time series of the latent variables and the observations has the following relationships

$$\begin{aligned} \mathbf{x}_i &= A\mathbf{x}_{i-1} + \mathbf{w}_i & \mathbf{w}_i &\sim N(\mathbf{w}_i | \mathbf{0}, \Gamma) \\ \mathbf{y}_i &= C\mathbf{x}_i + \mathbf{v}_i & \mathbf{v}_i &\sim N(\mathbf{v}_i | \mathbf{0}, \Sigma) \\ \mathbf{x}_1 &= \boldsymbol{\mu}_0 + \mathbf{u} & \mathbf{u} &\sim N(\mathbf{u} | \mathbf{0}, V_0), \end{aligned} \tag{1}$$

where $\mathbf{w}_i$, $\mathbf{v}_i$, $\mathbf{u}$ are noise terms. The LDS model fits the model parameters $\boldsymbol{\Theta} := \{A, C, \Gamma, \Sigma, \boldsymbol{\mu}_0, V_0\}$ by

taking the expectation over the latent variables $\mathbf{x}_i|\boldsymbol{\theta}_{old}$, and maximizing the log-likelihood of the complete data $\{\mathbf{x},\mathbf{y}|\boldsymbol{\theta}\}$, where $\boldsymbol{\theta}_{old}$ is the model parameter from the previous iteration and $\boldsymbol{\theta}$ is the parameter that we are seeking at the current iteration. In the expectation step, with the parameter $\boldsymbol{\theta}_{old}$, the mean and the variance of the posterior marginal latent variables $\mathbf{x}_i|\boldsymbol{\theta}_{old},\mathbf{y}_1,\mathbf{y}_2,\dots,\mathbf{y}_i$ at time i (see float 1 in Fig 7A) and the mean and the variance of the posterior marginal latent variable $\mathbf{x}_i|\boldsymbol{\theta}_{old},\mathbf{y}_1,\mathbf{y}_2,\dots,\mathbf{y}_T$ based on the information at all time (see float 1 in Fig 7B), are calculated using the forward and backward iterations. Here, T is the total length of a particular time series on a float. These marginal variables lead to the sufficient statistics of the complete data $\{\mathbf{x},\mathbf{y}|\boldsymbol{\theta}\}$, namely the $E_{\mathbf{x}|\boldsymbol{\theta}_{old},\mathbf{y}_1,\mathbf{y}_2,\dots,\mathbf{y}_T}[\mathbf{x}_i]$, $E_{\mathbf{x}|\boldsymbol{\theta}_{old},\mathbf{y}_1,\mathbf{y}_2,\dots,\mathbf{y}_T}[\mathbf{x}_i\mathbf{x}_i^T]$, $E_{\mathbf{x}|\boldsymbol{\theta}_{old},\mathbf{y}_1,\mathbf{y}_2,\dots,\mathbf{y}_T}[\mathbf{x}_i\mathbf{x}_i^T]$. Using these sufficient statistics [54, 55], we obtain the updated LDS model parameters $\boldsymbol{\Theta} = \{A, C, \Gamma, \Sigma, \boldsymbol{\mu}_0, V_0\}$. The graphical model [56] of the plain LDS is schematically plotted in Figs. 7A, 7B as one branch, for instance, the branch of float 1. The workflow of the plain LDS model is described in Algorithm 1.

Fig 7. (A) Information flow of the forward iteration in the Expectation step of the vLDS model for computing the mean and the variance of the posterior marginal latent variables $\mathbf{x}_i|\boldsymbol{\theta}_{old},\mathbf{y}_1,\mathbf{y}_2,\dots,\mathbf{y}_i$ at time i . **(B)** Information flow of the backward iteration in the Expectation step of the vLDS model for computing the mean and the variance of the posterior marginal latent variables $\mathbf{x}_i|\boldsymbol{\theta}_{old},\mathbf{y}_1,\mathbf{y}_2,\dots,\mathbf{y}_T$ based on information at all time $1\dots T$.

Algorithm 1. Plain version of training the LDS model on one float

1. initialize $\boldsymbol{\Theta} = \{A, C, \Gamma, \Sigma, \boldsymbol{\mu}_0, V_0\}$, $iter=1$, $maxiter=100$, $rtol = 10^{-4}$
2. **for** ($iter < maxiter$) **do**
3. Expectation step:
4. forward iteration to compute $\mathbf{x}_i|\boldsymbol{\theta},\mathbf{y}_1,\mathbf{y}_2,\dots,\mathbf{y}_i$
5. backward iteration to compute $\mathbf{x}_i|\boldsymbol{\theta},\mathbf{y}_1,\mathbf{y}_2,\dots,\mathbf{y}_T$

6. compute $E_{x|\theta, y_1, y_2, \dots, y_T}[x_i], E_{x|\theta, y_1, y_2, \dots, y_T}[x_i x_i^T], E_{x|\theta, y_1, y_2, \dots, y_T}[x_i x_i^T], llh(iter)$
 7. if $llh(iter) - llh(iter - 1) \geq rtol * llh(iter - 1)$ break; end if
 8. Maximization step:
 9. update $\Theta = \{A, C, \Gamma, \Sigma, \mu_0, V_0\}$
 10. end for
-

Variable-length Linear Dynamical Systems (vLDS)

In this study, each float generates one or more statistically independent time series of the *Chl a* concentration, due to the interpolation or splitting process discussed in the “Data Preprocessing” Section. For the preprocessed dataset with 186 floats, we treat it as multiple multivariate time series, each with a unique id. Also, we note that the lengths of the time series in the dataset are mostly different, due to the irregularity in the longevity of the floats. The variable-length Linear Dynamical Systems model is specifically designed for this situation, and it summarizes and recovers the latent dynamics from multiple multivariable time series with a different time span.

To fit the vLDS model, we start with some initial parameter Θ_0 , which is shared across all floats in the dataset. We keep the superscript n here. The Expectation step is carried out on each float id, using a two-loop forward and backward smoothing step to compute the conditional expectations of the sufficient statistics of the complete data $\{x, y | \theta\}$, namely the $E_{x|\theta_{old}, y_1, y_2, \dots, y_{T_n}}[x_i], E_{x|\theta_{old}, y_1, y_2, \dots, y_{T_n}}[x_i x_i^T], E_{x|\theta_{old}, y_1, y_2, \dots, y_{T_n}}[x_i x_{i-1}^T]$. In the maximization step, we use the averaging formula derived in Equation (5) across all floats to update the model parameter $\Theta = \{A, C, \Gamma, \Sigma, \mu_0, V_0\}$.

We emphasize that for a particular drifter n , the multivariate observation $y_i^n, i = 1, 2, 3, \dots, T_n$, contains all the physical and biological information along a drifter trajectory at time i , and it is the multivariate latent random variable x_i^n that we are solving for from the vLDS model to represent the hidden dynamics between different components of y_i^n , namely,

$\{lat, lon, ve, vn, spd, dist, chlor_a, par, cdm, t865, kd490, sst4\}$.

It is possible that some of the components of $\mathbf{y}_t^n$, for instance, the $t865$ in our study, as demonstrated in the “Discussion & Conclusion” Section, are not much involved in the latent dynamics. Therefore, the dimension of the latent space recovered by the latent variable $\mathbf{x}_t^n$ might be smaller than the dimension of the observations $\mathbf{y}_t^n$. In this study, the latent dimension in $\mathbf{x}_t^n$, as determined by the cross-validation procedure, turns out to be 11, and the dimension of the observations $\mathbf{y}_t^n$ is 12.

To test our benchmarked hypotheses in the “Background” Section, we first generate the predicted values of $\tilde{\mathbf{y}}_t^n$ by using the equation (1) with the recovered latent variables $\mathbf{x}_t^n$ from the vLDS model. Since the vLDS model automatically maximizing the log-likelihood of the complete data, including the latent variable $\mathbf{x}$ and the observational variable $\mathbf{y}$, it is beneficial to visualize the predictive plots of all predictors along the drifter trajectories, as shown in the “Discussion & Conclusion” Section, and study the performance metric R-squared, as described in the “Results” Section.

Probabilistic computation of vLDS

We assume that the observed multivariate variable $\mathbf{y}_t^n$, including the chlorophyll a concentration, of each float is evolving independently with other floats, except that they are driven by the same underlying biological and physical forces. It is this assumption that allows the vLDS model to share the same model parameters $\Theta = \{A, C, \Gamma, \Sigma, \mu_0, V_0\}$ in all branches in Fig 7A, 7B and makes the vLDS model a powerful algorithm to summarize and capture the population-level structures along drifter trajectories. Multiple variants of the LDS model have been applied successfully at the microscopic scale in several research areas such as computational neuroscience [19-20, 57-58] and sound tracking [22].

Algorithm 2. Training the vLDS model on the cross-validation dataset with many floats

```

1. initialize  $\Theta = \{A, C, \Gamma, \Sigma, \mu_0, V_0\}$ ,  $iter=1$ ,  $maxiter=100$ ,  $rtol = 10^{-4}$ 
2. for ( $iter < maxiter$ ) do
3.   Expectation step:
4.   for float  $n = 1, \dots, N$  do
5.     forward iteration to compute  $x_i | \theta, y_1, y_2, \dots, y_i$ 
6.     backward iteration to compute  $x_i | \theta, y_1, y_2, \dots, y_{T_n}$ 
7.     compute  $E_{x|\theta, y_1, y_2, \dots, y_{T_n}}[x_i]$ ,  $E_{x|\theta, y_1, y_2, \dots, y_{T_n}}[x_i x_i^T]$ ,  $E_{x|\theta, y_1, y_2, \dots, y_{T_n}}[x_i x_i^T]$ ,  $llh$ 
8.     sum  $llh(iter) = llh(iter) + llh$ 
9.   end for
10.  if  $llh(iter) - llh(iter - 1) \geq rtol \cdot llh(iter - 1)$  break; end if
11.  Maximization step:
12.  update  $\Theta = \{A, C, \Gamma, \Sigma, \mu_0, V_0\}$  using Equations (3) - (5)
13. end for

```

Although the distribution of the complete data $\{x, y | \theta\}$ depends on the model parameter θ , we omit the dependence on θ for notational ease in our following derivation. For a particular float n , letting i be the time step and T_n be the total length of the time series, the distribution of the complete data (namely the observations y and the latent variables x), can be written as

$$\begin{aligned}
 P(\{x^n, y^n\}) &= P(x_1^n) \prod_{i=2}^{T_n} P(x_i^n | x_{i-1}^n) \prod_{i=1}^{T_n} P(y_i^n | x_i^n) \\
 \log P(\{x^n, y^n\}) &= \log P(x_1^n) + \sum_{i=2}^{T_n} \log P(x_i^n | x_{i-1}^n) + \sum_{i=1}^{T_n} \log P(y_i^n | x_i^n) \\
 &= -\frac{1}{2} (x_1^n - \mu_0)^T V_0^{-1} (x_1^n - \mu_0) - \frac{1}{2} \\
 &\quad - \sum_{i=2}^{T_n} \left\{ \frac{1}{2} (x_i^n - A x_{i-1}^n)^T \Gamma^{-1} (x_i^n - A x_{i-1}^n) \right\} - \frac{T_n - 1}{2} \log |\Gamma| \\
 &\quad - \sum_{i=1}^{T_n} \left\{ \frac{1}{2} (y_i^n - C x_i^n)^T \Sigma^{-1} (y_i^n - C x_i^n) \right\} - \frac{T_n}{2} \log |\Sigma| + const.
 \end{aligned} \tag{2}$$

By the assumption of independence, the joint probability of the observations and state variables across all floats expands into the product of the joint probability of the observations and state

476 variables of all the time series generated by each float $n = 1, 2 \dots N$

$$477 \quad P(\{\mathbf{x}^1, \dots, \mathbf{x}^N, \mathbf{y}^1, \dots, \mathbf{y}^N\}) = \prod_{n=1}^N P(\mathbf{x}_1^n) \prod_{i=2}^{T_n} P(\mathbf{x}_i^n | \mathbf{x}_{i-1}^n) \prod_{i=1}^{T_n} P(\mathbf{y}_i^n | \mathbf{x}_i^n) \quad (3)$$

$$478 \quad \log P(\{\mathbf{x}^1, \dots, \mathbf{x}^N, \mathbf{y}^1, \dots, \mathbf{y}^N\}) = \sum_{n=1}^N \log P(\{\mathbf{x}^n, \mathbf{y}^n\}) \quad (4)$$

479

480 We note that, for each float, the preprocessed data of this float might generate multiple
 481 multivariate time series (see the “Data Preprocessing” Section for more details.) Under the
 482 traditional i.i.d. assumptions, the objective function of the vLDS model is simply the addition of the
 483 objective functions for each individual time series' log-likelihood with all the parameters $\Theta :=$
 484 $\{A, C, \Gamma, \Sigma, \mu_0, V_0\}$ that define the plain version LDS for each float (or each box-branch in Fig 7) being
 485 shared across all the floats. Using the Expectation-Maximizing algorithm, the Expectation step can
 486 be carried out using a two loop backward and forward iteration for each time series independently,
 487 due to the conditional independence of the state variables across different time series of different
 488 floats. However, in the maximization step we need to average the vLDS model parameter across all
 489 the time series from all the floats.

490

491 The derivation of the update formula for the vLDS model parameters $\Theta := \{A, C, \Gamma, \Sigma, \mu_0, V_0\}$ follows
 492 directly by taking derivatives of the complete data log-likelihood with respect to each component
 493 in Θ and by using the standard results from the Maximum Likelihood Estimators of the mean and
 494 variance for the Gaussian Distribution. For a dataset of many floats, the complete data log-likelihood
 495 has an addition summation sign running through $n = 1, 2 \dots N$ in Equation (4). We use T_n , instead of
 496 T , to denote the length of the time series of the float n . Given the fact that the derivative of a linear
 497 combination of functions is a linear combination of derivatives of each function, we can write the
 498 updating formula for the maximization step as:

$$\begin{aligned}
 C^{new} &= \left(\sum_{n=1}^N \sum_{i=1}^{T_n} \mathbf{y}_i^n E(\mathbf{x}_i^n)^T \right) \left(\sum_{n=1}^N \sum_{i=1}^{T_n} E(\mathbf{x}_i^n (\mathbf{x}_i^n)^T) \right)^{-1} \\
 \Sigma^{new} &= \frac{1}{\sum_{n=1}^N T_n} \left(\sum_{n=1}^N \sum_{i=1}^{T_n} (\mathbf{y}_i^n (\mathbf{y}_i^n)^T - C^{new} E(\mathbf{x}_i^n) (\mathbf{y}_i^n)^T - \mathbf{y}_i^n E(\mathbf{x}_i^n)^T C^{new} \right. \\
 &\quad \left. + C^{new} E(\mathbf{x}_i^n (\mathbf{x}_i^n)^T) C^{new} \right) \\
 A^{new} &= \left(\sum_{n=1}^N \sum_{i=2}^{T_n} E(\mathbf{x}_i^n (\mathbf{x}_{i-1}^n)^T) \right) \left(\sum_{n=1}^N \sum_{i=2}^{T_n} E(\mathbf{x}_{i-1}^n (\mathbf{x}_{i-1}^n)^T) \right)^{-1} \quad (5) \\
 \Gamma^{new} &= \frac{1}{\sum_{n=1}^N (T_n - 1)} \left(\sum_{n=1}^N \sum_{i=2}^{T_n} (E(\mathbf{x}_i^n (\mathbf{x}_i^n)^T - A^{new} E(\mathbf{x}_{i-1}^n (\mathbf{x}_i^n)^T) - E(\mathbf{x}_i^n (\mathbf{x}_{i-1}^n)^T) A^{new} \right. \\
 &\quad \left. + A^{new} E(\mathbf{x}_{i-1}^n (\mathbf{x}_{i-1}^n)^T) (A^{new})^T) \right) \\
 \boldsymbol{\mu}_0^{new} &= \frac{1}{N} \sum_{n=1}^N E(\mathbf{x}_1^n) \\
 \mathbf{V}_0^{new} &= \frac{1}{N} \sum_{n=1}^N (E(\mathbf{x}_1^n (\mathbf{x}_1^n)^T) - E(\mathbf{x}_1^n) E((\mathbf{x}_1^n)^T))
 \end{aligned}$$

The workflow of the vLDS model is described in Algorithm 2, and the averaging of the vLDS model parameters Θ across all the floats in equation (4) is carried out in step 12. During this procedure, the vLDS model adaptively learns the latent dynamics of the underlying process. We have fitted the chlorophyll *a* concentration for the particular float in Fig 5 with id 64113560. The result is displayed in the “Discussion and Conclusion” Section.

Each Expectation-Maximization cycle of the LDS model for Gaussian random variables is guaranteed to increase the value of the complete data log-likelihood. Therefore, a standard stopping

criteria for the Expectation-Maximization algorithm is based on the complete data log-likelihood in Equations (2), (4) with a relative tolerance $rtol = 10^{-4}$ and maximum iteration 100.

One of the key model parameters in the LDS modeling is the dimension k of the latent space, namely, the number of components in the latent variable $\mathbf{x}$. It is the dimension of the subspace generated by the projection of the full feature space onto the latent subspace, whose projection back onto the full feature space in Equation (1) under the vLDS linear transformation matrix C maximizes the complete data log-likelihood. A larger k indicates that there are more independent factors in the latent space of $\mathbf{x}$ driving the underlying dynamical system of $\{\mathbf{x}, \mathbf{y}\}$. Moreover, varying the values of the dimensionality k induces a family of different vLDS models (1) - (3) indexed by k . To select the model with the most appropriate parameter k , we carry out a 10-fold cross-validation [59-61] on the parameter k and choose the optimal k that achieves the maximum complete data log-likelihood on the test dataset. More specifically, we group the dataset by float ids. We hold a portion of the float ids and consider them as the heldout dataset. We take the rest of the float ids as the cross-validation dataset. In the cross-validation step, we split the cross-validation set evenly into 10 folds. Each time we take one fold as the testing dataset, we take the rest as the training dataset. We fit the vLDS parameter $\Theta = \{A, C, \Gamma, \Sigma, \mu_0, V_0\}$ on the training dataset and compute the complete data log-likelihood on the testing dataset using this newly fitted parameter Θ . The complete data log-likelihood is averaged for different testing fold for a fixed k . Then, we repeat the process for different values of k . See Table 2 and Fig 8 for the complete data log-likelihood generated by different cross-validation trails. The averaged complete-data log-likelihood across different testing set is maximized by $k = 11$.

Table 2. 10-fold cross-validation on the parameter k . The optimal k that achieves the maximum complete data log-likelihood on the test set is $k = 11$. The unit of the test-dataset log-likelihood in the table is 10^4 .

	k	1	2	3	4	5	6	7	8	9	10	11	12
fold	llh_{test}												
1		-1.44	-1.36	-1.33	-1.23	-1.15	-1.19	-1.15	-1.04	-1.14	-1.18	-1.01	-1.65
2		-1.33	-1.22	-1.20	-1.09	-0.98	-1.02	-0.95	-0.83	-0.94	-0.97	-0.81	-1.42
3		-1.58	-1.53	-1.47	-1.27	-1.25	-1.19	-1.26	-1.10	-1.19	-1.15	-1.02	-1.46
4		-1.29	-1.18	-1.15	-1.04	-0.98	-0.97	-0.94	-0.82	-0.88	-0.86	-0.80	-1.11
5		-1.44	-1.33	-1.30	-1.17	-1.12	-1.11	-1.10	-0.91	-1.06	-1.11	-0.88	-1.57
6		-1.58	-1.45	-1.45	-1.34	-1.22	-1.36	-1.31	-1.17	-1.27	-1.19	-1.14	-1.70
7		-1.71	-1.55	-1.52	-1.34	-1.25	-1.22	-1.19	-1.02	-1.07	-1.09	-0.95	-1.31
8		-1.14	-1.02	-1.02	-0.93	-0.84	-0.88	-0.87	-0.73	-0.76	-0.76	-0.71	-0.95
9		-1.63	-1.48	-1.44	-1.27	-1.17	-1.18	-1.19	-0.99	-1.06	-1.10	-0.94	-1.35
10		-1.52	-1.40	-1.35	-1.22	-1.20	-1.13	-1.11	-0.92	-0.99	-1.04	-0.85	-1.33
Average		-1.47	-1.35	-1.32	-1.19	-1.11	-1.12	-1.11	-0.95	-1.04	-1.05	-0.92	-1.38

Fig 8. 10-fold cross-validation on the parameter k and the optimal k that achieves the maximum complete data log-likelihood on the test set is $k = 11$.

With the optimal value $k = 11$ of the latent space dimension identified, we fit the vLDS model one more time with the full cross-validation dataset to generate the vLDS model parameter. In Fig 9, we display the log-likelihood convergence of the Expectation-Maximization algorithm for the complete cross-validation dataset and five individual floats in the cross-validation dataset. We note that from Equations (2) - (4), the log-likelihood of the complete cross-validation dataset is the sum of the log-likelihood of each individual floats in the cross-validation dataset (step 8 in Algorithm 2).

Fig 9. Convergence history of the log-likelihood of the complete cross-validation dataset and a sample of the convergence history for 5 floats.

Results

With the optimally selected latent space dimension $= 11$, the vLDS algorithm obtains a set of model parameters $\Theta = \{A, C, \Gamma, \Sigma, \mu_0, V_0\}$ when the stopping criteria inside the Expectation-Maximization algorithm is reached. In Fig 10, we display the prediction results for some drifter ids in the cross-validation dataset, using Equation (1) and the expected conditional mean of the latent variables at the last iteration of the Expectation steps 4, 5, and 6 in Algorithm 2. The dark lines are the observations, and the cyan lines are the predictions. Most of the hidden dynamics of the float profiles inside the cross-validation dataset are well captured by the vLDS model. We note the positive correlations among ‘*chlor_a*’, ‘*cdm*’, and ‘*kd490*’ in the recovered vLDS latent dynamics (cyan lines in Fig 10) at the microscopic population-level for all drifters in the cross-validation dataset. The model captures this correlation with some overshooting or undershooting in certain regions. Also, ‘*t865*’, the aerosol optical thickness over water, turns out to be independent of the chlorophyll *a* concentration and other ocean profiles. Moreover, the spatial information, namely, the longitude, latitude, velocity, speed of the float, and distance from the nearest coast, is all well recovered by the vLDS model (*lat* and *spd* are not shown in Figs 10, 11 due to space limitations).

We next examine on the robustness of the vLDS model. The floats in the heldout dataset are not used in the cross-validation process or the model’s parameter estimation process. Therefore, the heldout dataset is totally unknown to the vLDS learning algorithm. We use the cross-validated latent dimension $k = 11$, and the model parameter $\Theta = \{A, C, \Gamma, \Sigma, \mu_0, V_0\}$ generated by training the vLDS

584 model on the cross-validation dataset. Applying one iteration of the forward-backward smoothing
585 process, namely, merely one iteration of the Expectation steps 4, 5, and 6 in Algorithm 2, to each
586 float in the heldout dataset, we obtain the predictions of their profiles (Fig 11). Most of the hidden
587 dynamics along drifter trajectories for the floats in the heldout dataset, which is totally

588

589

590 **Fig 10.** Predictions of the drifter profiles for the floats in the cross-validation dataset, using the expected conditional mean of the latent
591 variables at the last iteration of the Expectation-Maximization algorithm.

592

593

594

595 **Fig 11.** Predictions of the drifter profiles for the floats in the heldout dataset, using the expected conditional mean of the latent variables
596 generated by one iteration of the forward-backward smoothing process of the Expectation-Maximization algorithm with the vLDS
597 model parameter $\Theta = \{A, C, \Gamma, \Sigma, \mu_0, V_0\}$ optimized on the cross-validated dataset.

unknown to the learning algorithm, is well captured by the vLDS model. It clearly demonstrates the generalizability of the vLDS model's capability to summarize and capture the population-level structures along the drifter trajectories on unknown datasets.

We note again the positive correlations among '*chlor_a*', '*cdm*', and '*kd490*' in the recovered vLDS latent dynamics (cyan lines in Fig 11) at the microscopic population-level for the drifters in the heldout dataset. Even in this heldout dataset, the model is capable of capturing this correlation with some overshooting or undershooting in certain regions. Again, '*t865*', the aerosol optical thickness over water, seems to be independent of the *Chl a* concentration and other ocean profiles in the heldout dataset. The vLDS model simply estimate the mean value for the variable '*t865*' in both the cross-validation and heldout datasets. However, by comparing the predictions generated by the vLDS model for '*t865*' and for other variables, we conclude that the variable '*t865*' is not much involved in the latent dynamics of the *Noctiluca*'s growth. Otherwise, the predicted values of '*t865*' should match its observations in Fig 10, 11, and the R^2 of '*t865*' in Table 3 should not be too small. Evidently, the vLDS model provides us the opportunity to verify the causal relationships of all variables at the microscopic population-level along the drifter trajectories, a feature the must be lacking in any models that do not take the trajectory-based population-level structure into consideration. Moreover, the spatial information of the heldout floats, namely, longitude, latitude, velocity, speed of the float, and distance from the nearest coast, is all well recovered by the vLDS model.

To quantify the performance of the vLDS model, we use the R-squared metric (R^2). In Table 3, the total sum of squares (SSTotal), sum of squared errors of predictions (SSE), and R^2 , which is the portion of the variance explained by the predictive model, are computed for both the cross-

validation and heldout datasets. Although the vLDS model has the log-likelihood of the complete data in Equations (3), (4) as its own performance metric, we use R^2 here for an intuitive interpretation. The quantitative results reflect the visualization in Fig 10, 11. The feature 't865' has a very small value in its R^2 metric and does not exhibit any predictive power. The vLDS recovers the spatio-temporal information well, and explains most of the variance in the biological factors {'cdm', 'kd490', 'par', 'sst4', 'chlor_a'}.

Table 3. R-squared (R^2) metric for the cross-validation dataset and heldout dataset. The R^2 is computed for each individual feature and for all the features aggregated together.

	cross-validation data			heldout data		
	SSTotal	SSE	R^2	SSTotal	SSE	R^2
<i>lat</i>	1.17E+05	3.78E+02	0.99	4.32E+03	1.86E+01	0.99
<i>lon</i>	2.32E+05	4.42E+02	0.99	3.52E+03	3.23E+01	0.99
<i>ve</i>	2.37E+06	3.67E+05	0.85	1.51E+05	2.49E+04	0.84
<i>vn</i>	2.36E+06	4.51E+05	0.81	1.07E+05	1.63E+04	0.85
<i>spd</i>	1.78E+06	2.62E+05	0.85	7.20E+04	1.09E+04	0.85
<i>dist</i>	3.55E+08	2.05E+06	0.99	1.78E+07	1.52E+05	0.99
<i>cdm</i>	1.65E+01	1.88E+00	0.89	3.40E-01	4.00E-02	0.87
<i>kd490</i>	2.43E+01	8.02E+00	0.67	3.19E-01	1.39E-01	0.56
<i>t865</i>	1.34E+01	1.29E+01	0.04	7.30E-01	7.49E-01	-0.04
<i>par</i>	2.41E+05	1.80E+04	0.93	9.43E+03	7.87E+02	0.91
<i>sst4</i>	1.20E+04	2.07E+03	0.83	1.02E+03	5.13E+02	0.49
<i>chlor_a</i>	4.60E+04	2.19E+04	0.52	2.59E+02	1.46E+02	0.43
Aggregated	3.62E+08	3.17E+06	0.99	1.82E+07	2.06E+05	0.98

Discussion & conclusion

We have introduced a new model vLDS and showed it offers a new approach to recover microscopic biogeochemical mechanisms underlying chaotic drifter trajectories that might be unobservable at macroscopic scale or accessible only in controlled laboratory experiments. The vLDS model generates predictions that recover the causal relationship among the *Noctiluca* blooms, physical dispersal, and biological environments (Fig 10, 11, Table 3.) The model's generalization capability

also summarizes, recovers, and predicts the latent dynamics from unknown heldout datasets, thus inspiring confidence in our microscopic and macroscopic findings. The highly correlated relationships between the ‘*chlor_a*’ and ‘*cdm*’ (colored dissolved organic matter *CDOM*), and between the ‘*chlor_a*’ and ‘*kd490*’ (light under the sea surface) are close to linear. The tightly correlated relationships between the ‘*chlor_a*’ and ‘*par*’ (light on the sea surface *PAR*), and between the ‘*chlor_a*’ and ‘*sst4*’ (sea surface temperature *SST4*) are nonlinear. The atmospheric deposition ‘*t865*’ is not involved in the underlying dynamics of the *Noctiluca* blooms.

These results confirm our hypotheses in the “Background” Section from the microscopic perspective, and confirms the impact of both the physical transport and biological factors of light, and nutrients, proxies for the latter being *CDOM* [26, 62-65], on the distribution of the *Noctiluca* blooms. Also, from the test results, the nutrient and light are two important positive factors for the *Noctiluca*’s growth and that the atmospheric deposition does not contribute much to this process. More specifically, the nutrient-rich water flowing into sea are increasing productivity and expanding the hypoxia zone as oxygen gets depleted. Since the atmospheric deposition measured by *T865* does not contribute much to the source of nutrient (Fig 10, 11, Table 3), the potential sources of the organic matter are including the domestic and industrial outfall from countries bordering the Arabian Sea, and the Northeast Monsoon (NEM, Nov–Jan) and winter convective mixing of deep nutrient rich, oxygen poor waters. As plankton grows, it uses up oxygen in the water. Falling oxygen levels allow *Noctiluca* to outcompete diatoms during winter. It has also been confirmed from the microscopic dynamics recovered by the vLDS Model that *Noctiluca* grow faster in lighted than in dark areas on the sea surface and in the sea water.

We have demonstrated the effectiveness of the vLDS model as a microscopic statistical modeling approach for detecting important causal relationships in biogeochemical processes. Although the trajectories of the oceanographic probing devices are chaotic and the dataset is high dimensional, the vLDS model is very parsimonious on model parameters. The model only requires $\Theta = \{A, C, \Gamma, \Sigma, \mu_0, V_0\}$ and the latent-space dimension k to be able to summarize all the drifters in the Arabian Sea region from 2002 to 2017. The predictive dynamics matches the microscopic observations well, and provides us with tremendous confidence to support our macroscopic hypotheses.

Furthermore, the intertwined relationships recovered by the vLDS model between the physical and biological dynamics of the *Noctiluca* blooms, and the intertwined relationships among the biological factors such as '*chlor_a*', '*cdm*', and '*kd490*' have inspired us to use inference tools to quantify the isolated impact of the biological factors that are responsible for the *Noctiluca* blooms in the Arabian Sea region. The vLDS model presented here is fully generalizable to other datasets for applications in other fields i.e. larval transport etc. in marine larval ecology.

Acknowledgements

The authors gratefully acknowledge the support from the OceanColor Team, GlobColor Team and Xarray Development Team during the data collection and preprocessing process.

References

1. Leathwick JR, Elith J, Francis MP, Hastie T, Taylor P. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. Marine Ecology Progress Series. 2006;321: 267-281.

2. Murray KT, Orphanides CD. Estimating the risk of loggerhead turtle *Caretta caretta* bycatch in the US mid-Atlantic using fishery-independent and -dependent data. *Mar Ecol Prog Ser.* 2013;477: 259-270.
3. Rooper CN, Zimmermann M, Prescott MM, Hermann AJ. Predictive models of coral and sponge distribution, abundance and diversity in bottom trawl surveys of the Aleutian Islands, Alaska. *Mar Ecol Prog Ser.* 2014;503: 157-176.
4. Schmiing M, Afonso P, Tempera F, Santos RS. Predictive habitat modelling of reef fishes with contrasting trophic ecologies. *Mar Ecol Prog Ser.* 2013;474: 201-216.
5. Windle MJS, Rose GA, Devillers R, Fortin MJ. Spatio-temporal variations in invertebrate-cod-environment relationships on the Newfoundland-Labrador Shelf, 1995–2009. *Mar Ecol Prog Ser.* 2012;469: 263-278.
6. Zanardo N, Parra GJ, Passadore C, Möller LM. Ensemble modelling of southern Australian bottlenose dolphin *Tursiops* sp. distribution reveals important habitats and their potential ecological function. *Mar Ecol Prog Ser.* 2017;569: 253-266.
7. Lima C, Lall U, Jebara T, Barnston AG. Statistical prediction of ENSO from subsurface sea temperature using a nonlinear dimensionality reduction. *Journal of Climate.* 2009;22(17): 4501-4519.
8. Castellari S, Griffa A, Ozgokmen TM, Poulain PM. Prediction of particle trajectories in the Adriatic Sea using Lagrangian data assimilation. *Journal of Marine Systems.* 2001;29(1-4): 33-50.
9. Chin TM, Mariano AJ. A particle filter for inverse Lagrangian prediction problems. *Journal of Atmospheric and Oceanic Technology.* 2010;27(2): 371-384.
10. Bengtsson T, Milliff R, Jones R, Nychka D, Niiler PP. A state-space model for ocean drifter motions dominated by inertial oscillations. *Journal of Geophysical Research: Oceans.* 2005;110: C10015.
11. Slivinski L, Spiller E, Apte A, Sandstede B. A hybrid particle-ensemble Kalman filter for Lagrangian data assimilation. *Mon Wea Rev.* 2015;143: 195-211.
12. Sameh A, Ltaief H, Sun Y, Genton MG, Keyes DE. ExaGeoStat: A high performance unified framework for Geostatistics on manycore systems. Preprint. Available from: arXiv:1708.02835, 2017.
13. Tang D, Jebara T. Initialization and coordinate optimization for multi-way matching. *Artificial Intelligence and Statistics (AISTATS)*, 2017.
14. Shaw B, Jebara T. Structure preserving embedding. *International Conference on Machine Learning (ICML)*, 2009.

15. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2009;31(2): 210-227.
16. Kerman J, Gelman A, Zheng T, Ding Y. Visualization in Bayesian data analysis. In: *Handbook of data visualization*. Springer Handbooks Comp Statistics. Springer, Berlin, Heidelberg; 2008.
17. Merel J, Fox R, Jebara T, Paninski L. A multi-agent control framework for co-adaptation in brain-computer interfaces. *Neural Information Processing Systems (NIPS)*, 2013.
18. Pakman A, Huggins J, Smith C, and Paninski L. Fast state-space methods for inferring dendritic synaptic connectivity. *Journal of Computational Neuroscience*. 2014;36(3): 415-43.
19. Elsayed GF, Cunningham JP. Structure in neural population recordings: an expected byproduct of simpler phenomena? *Nature Neuroscience*. 2017;20: 1310-1318.
20. Linderman S, Miller A, Adams R, Blei D, Paninski L, Johnson M. Recurrent switching linear dynamical systems. *Artificial Intelligence and Statistics (AISTATS)*, 2017.
21. Rosenthal S, McKeown K. Columbia nlp: Sentiment detection of subjective phrases in social media. In: *Proceedings of the 7th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Atlanta, Georgia; 2013.
22. Benetos E, Lafay G, Lagrange M, Plumbley MD. Polyphonic sound event tracking using linear dynamical systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 2017;25(6): 1266-1277.
23. Rabinovich M, Blei D. The inverse regression topic model. *International Conference on Machine Learning (ICML)*, 2014.
24. Kandula S, Hsu D, Shaman J. Subregional nowcasts of seasonal influenza using search trends. *J Med Internet Res*. 2017;19(11): e370.
25. Gomes H do R, Goes JI, Matondkar SGP, Parab SG, Al-Azri ARN, Thoppil PG. Blooms of noctiluca miliaris in the Arabian Sea - an in situ and satellite study. *Deep Sea Research Part I: Oceanographic Research Papers*. 2008;55(6): 751-765.
26. Gomes H do R, Goes JI, Matondkar SGP, Buskey EJ, Basu S, Parab SG, Thoppil P. Massive outbreaks of noctiluca scintillans blooms in the Arabian Sea due to spread of hypoxia. *Nature Communications*. 2014;5: 4862.
27. Goes JI, Gomes H do R, Al-Azri ARN, Al-Hashmi K editors. An assessment of environmental drivers responsible for the emergence of mixotrophy in the Arabian Sea. *Proceedings of the 2016 Ocean Sciences Meeting*; 2016 Feb; American Geophysical Union.

28. Goes JI, Gomes H do R, Al-Hashimi K, Buranapratheprat A. Ecological drivers of green *Noctiluca* blooms in two monsoonal-driven ecosystems. In: Glibert P, Berdalet E, Burford M, Pitcher G, Zhou M, editors. *Global Ecology and Oceanography of Harmful Algal Blooms. Ecological Studies (Analysis and Synthesis)*, vol 232. Springer, Cham; 2018.
29. Margalef R. Life forms of phytoplankton as survival alternatives in an unstable environment. *Oceanology Acta*. 1978;1: 493-509.
30. Reynolds CS. Community organization in the freshwater plankton. *Symp Br Ecol Soc*. 1987;27: 297-325.
31. Diehl S, Berger S, Ptacnik R, Wild A. Phytoplankton, light, and nutrients in a gradient of mixing depths: Field Experiments. *Ecology*. 2002;83(2): 399-411.
32. Huisman J, Sharples J, Stroom JM, Visser PM, Kardinaal WE, Verspagen JM, Sommeijer B. Changes in turbulent mixing shift competition for light between phytoplankton species. *Ecology*. 2004;85: 2960-2970.
33. Gomes H do R, Xu Q, Ishizaka J, Carpenter EJ, Yager PL, Goes JI. The influence of riverine nutrients in niche partitioning of phytoplankton communities – a contrast between the Amazon River Plume and the ChangJiang (Yangtze) River diluted water of the East China Sea. *Front Mar Sci*. 2018; Forthcoming. doi: 10.3389/fmars.2018.00343.
34. Abernathey R, Ferreira D, Klocker A. Diagnostics of isopycnal mixing in a circumpolar channel. *Ocean Modelling*. 2013;72: 1-16.
35. Abernathey R, Marshall J. Global surface eddy diffusivities derived from satellite altimetry. *Journal of Geophysical Research: Oceans*. 2013;118(2): 901-916.
36. Chelton DB, Gaube P, Schlax MG, Early JJ, Samelson RM. The influence of nonlinear mesoscale eddies on near-surface oceanic chlorophyll. *Science*. 2011;334(6054): 328-332.
37. Klocker A, Abernathey R. Global patterns of mesoscale eddy properties and diffusivities. *Journal of Physical Oceanography*. 2014;44(3): 1030-1046.
38. Lévy M, Ferrari R, Franks PJS, Martin AP, Rivire P. Bringing physics to life at the submesoscale. *Geophysical Research Letters*. 2012;39(14): L14602.
39. Thomas LN, Tandon A, Mahadevan A. Sub-mesoscale processes and dynamics. In: Hecht MW, Hasumi H, editors. *Ocean modeling in an eddying regime. Geophysical Monograph Series, Volume 177*. American Geophysical Union, Washington DC; 2008. p. 17-38.
40. Goes JI, Thoppil PG, Gomes H do R, Fasullo JT. Warming of the Eurasian landmass is making the Arabian Sea more productive. *Science*. 2005;308(5721): 545-547.
41. McGillicuddy DJ, Anderson LA, Bates NR, Bibby T, Buesseler KO, Carlson CA, et al. Eddy/wind interactions stimulate extraordinary mid-ocean plankton blooms. *Science*. 2007;316(5827): 1021-1026.

42. OceanColor, 2016. Available from: <http://oceancolor.gsfc.nasa.gov> (accessed 06 Jun 2016).
43. Lumpkin R, Pazos M. Measuring surface currents with Surface Velocity Program drifters: the instrument, its data, and some recent results. In: Griffa A, Kirwan AD, Mariano AJ, Ozgokmen T, Rossby T, editors. Chapter two of Lagrangian analysis and prediction of coastal and ocean dynamics (LAPCOD). Cambridge University Press; 2007. p. 39-67.
44. NASA Goddard Space Flight Center, Ocean Ecology Laboratory, Ocean Biology Processing Group. Moderate-resolution imaging spectroradiometer (MODIS) Aqua photosynthetically available radiation data; 2014 Reprocessing. NASA OB.DAAC, Greenbelt, MD, USA. doi:10.5067/AQUA/MODIS/L3B/PAR/2014 (accessed 05 Jun 2017).
45. GlobColour data used in this study has been developed, validated, and distributed by ACRI-ST, France, 2017. Available from: <http://globcolour.info> (accessed 07 Jun 2017).
46. Maritorena S, Hembise Fanton d'Andon O, Mangin A, Siegel DA. Merged satellite ocean color data products using a bio-optical model: Characteristics, benefits and issues. *Remote Sensing of Environment*. 2010;114(8): 1791-1804.
47. Fanton d'Andon O, Mangin A, Lavender S, Antoine D, Maritorena S, Morel A, et al. GlobColour - the European service for Ocean Colour. In: Proceedings of the 2009 IEEE International Geoscience & Remote Sensing Symposium, Jul 12-17 2009, Cape Town South Africa. IEEE Geoscience and Remote Sensing Society; 2009.
48. Moisan JR, Niiler PP, Abbott M, Letelier R. The use of Lagrangian drifters to measure biogeochemical processes and to analyze satellite data sets. In: Harada K, Dickey T, editors. Proceedings of the International Workshop on Autonomous Measurements of Biogeochemical Parameters in the Ocean. Ocean Phys Lab. University of California, Santa Barbara, CA. 2001.
49. Aref H. Stirring by chaotic advection. *J Fluid Mech*. 1984;143: 1-21.
50. Samelson RM. Chaotic transport by mesoscale motions. In: Aller RJ, Miller P, Rozovskii B, editors. *Stochastic Modeling in Physical Oceanography*. Birkhauser; 1996. p. 423-438.
51. Abbott MR, Letelier RM. De-correlation scales of chlorophyll as observed from bio-optical drifters in the California Current. *Deep Sea Res Part II*. 1998;45: 1639-1667.
52. Veneziani M, Griffa A, Poulain PM. Historical drifter data and statistical prediction of particle motion: A case study in the central Adriatic Sea. *Journal of Atmospheric and Oceanic Technology*. 2007;24(2): 235-254.
53. Mahadevan A. Spatial heterogeneity and its relation to processes in the upper ocean. In: Lovett GM, Turner MG, Jones CG, Weathers KC, editors. *Ecosystem function in heterogeneous landscapes*. Springer New York, New York, NY; 2005. P. 165-182.
54. Bishop CM. *Pattern recognition and machine learning*. Springer Verlag New York, NY; 2006.

55. Ghahramani Z, Hinton G. Parameter estimation for linear dynamical systems. Technical Report CRG-TR-96-2, Department of Computer Science, University of Toronto; 1996.
56. Jordan MI. Graphical models. *Statistical Science. Special Issue on Bayesian Statistics.* 2004;19: 140-155.
57. Gao Y, Archer E, Paninski L, Cunningham JP. Linear dynamical neural population models through nonlinear embeddings. NIPS, 2016.
58. Gao Y, Buesing L, Shenoy KV, Cunningham JP. High-dimensional neural spike train analysis with generalized count linear dynamical systems. NIPS, 2015.
59. Breiman L, Spector P. Submodel selection and evaluation in regression: The X-random case. *International Statistical Review.* 1992;60: 291-319.
60. Zhang P. Model selection via multifold cross-validation. *Annals of Statistics.* 1993;21: 299-311.
61. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2 (IJCAI'95).* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA; 1995. p. 1137-1143.
62. Dhage SS, Chandorkar AA, Kumar R, Srivastava A, Gupta I. Marine water quality assessment at Mumbai West Coast. *Environment International.* 2006;32(2): 149-158.
63. Murthy RC, Rao YR, Inamdar AB. Integrated coastal management of Mumbai metropolitan region. *Ocean & Coastal Management.* 2001;44(5–6): 355-369.
64. Hussain M. Competitive Performance Evaluation of Waste Water Treatment Plants of Karachi and Impact of Untreated Waste Water on Some Edible Fishes of Arabian Sea. Ph.D. Thesis, University of Karachi. 2007.
65. Dang GT. Variability in the Colored and Fluorescent Dissolved Organic Matter with Salinity in the Neuse River Estuary. M.Sc. Thesis, The University of North Carolina at Chapel Hill. 2011.

(A)

(B)

(C)

Temperature trajectories of all drifters in the Arabian Sea Region from 2002 to 2017

id - 64113560

10

id - 64113560

id - 71140

id - 71140.06

id - 71140.15

vLDS parameter
 $\Theta := \{A, C, \Gamma, \Sigma, \mu_0, V_0\}$

(A)

vLDS parameter
 $\Theta := \{A, C, \Gamma, \Sigma, \mu_0, V_0\}$

(B)

maximum log-likelihood on the test dataset in the 10-fold cross-validation
is achieved at latent dimension $k=11$

log-likelihood at each iteration on the cross-validation dataset
using 10-fold cross-validated latent dimension $k=11$

— observations of cv data
— predictions of cv data

— observations of heldout data
 — predictions of heldout data