



HAL
open science

Application of Machine Learning Techniques to Ocean Mooring Time Series Data

Bernadette Sloyan, Christopher Chapman, Rebecca Cowley, Anastase Alexandre Charantonis

► **To cite this version:**

Bernadette Sloyan, Christopher Chapman, Rebecca Cowley, Anastase Alexandre Charantonis. Application of Machine Learning Techniques to Ocean Mooring Time Series Data. *Journal of Atmospheric and Oceanic Technology*, 2023, 40 (3), pp.241-260. 10.1175/jtech-d-21-0183.1 . hal-04104564

HAL Id: hal-04104564

<https://hal.science/hal-04104564v1>

Submitted on 25 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Application of Machine Learning Techniques to Ocean Mooring Time Series Data

BERNADETTE M. SLOYAN,^a CHRISTOPHER C. CHAPMAN,^a REBECCA COWLEY,^a AND ANASTASE A. CHARANTONIS^{b,c,d}

^a Centre for Southern Hemisphere Oceans Research, Oceans and Atmosphere, CSIRO, Hobart, Tasmania, Australia

^b Laboratoire d'Océanographie et du Climat Expérimentations et Approches Numériques, Sorbonne Université, Paris, France

^c École Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise, Évry, France

^d Laboratoire de Mathématiques et Modélisation d'Évry, Évry, France

(Manuscript received 3 January 2022, in final form 8 December 2022)

ABSTRACT: In situ observations are vital to improving our understanding of the variability and dynamics of the ocean. A critical component of the ocean circulation is the strong, narrow, and highly variable western boundary currents. Ocean moorings that extend from the seafloor to the surface remain the most effective and efficient method to fully observe these currents. For various reasons, mooring instruments may not provide continuous records. Here we assess the application of the Iterative Completion Self-Organizing Maps (ITCOMPSON) machine learning technique to fill observational data gaps in a 7.5 yr time series of the East Australian Current. The method was validated by withholding parts of fully known profiles, and reconstructing them. For 20% random withholding of known velocity data, validation statistics of the u - and v -velocity components are R^2 coefficients of 0.70 and 0.88 and root-mean-square errors of 0.038 and 0.064 m s⁻¹, respectively. Withholding 100 days of known velocity profiles over a depth range between 60 and 700 m has mean profile residual differences between true and predicted u and v velocity of 0.009 and 0.02 m s⁻¹, respectively. The ITCOMPSON also reproduces the known velocity variability. For 20% withholding of salinity and temperature data, root-mean-square errors of 0.04 and 0.38°C, respectively, are obtained. The ITCOMPSON validation statistics are significantly better than those obtained when standard data filling methods are used. We suggest that machine learning techniques can be an appropriate method to fill missing data and enable production of observational-derived data products.


SIGNIFICANCE STATEMENT: Moored observational time series of ocean boundary currents monitor the full-depth variability and change of these dynamic currents and are used to understand their influence on large-scale ocean climate, regional shelf-coastal processes, extreme weather, and seasonal climate. In this study we apply a machine learning technique, Iterative Completion Self-Organizing Maps (ITCOMPSON), to fill data gaps in a boundary current moored observational data record. The ITCOMPSON provides an improved method to fill data gaps in the mooring record and if applied to other observational data records may improve the reconstruction of missing data. The derived gridded data product should improve the accessibility and potentially increase the use of these data.

KEYWORDS: Currents; Ocean circulation; In situ oceanic observations; Neural networks

1. Introduction

Ocean boundary currents are narrow, strong flows that move large amounts of water, heat, and other ocean properties within the global ocean (Pedlosky 1996; Talley et al. 2011). They are an important component of the ocean circulation and strongly influence ocean and climate variability. A particular focus of study are the western boundary currents (WBCs) due to their dominant role in balancing the large-scale wind driven basin gyres and redistribution of heat and ocean properties from the low latitudes to the mid- and high latitudes (Todd et al. 2019). WBCs, in addition, also respond to local forcing variability such as wind and buoyancy. Thus, WBCs characteristics and property transports are driven by both large-scale and local forcing that occur over time scales of days to decades.

WBCs are found at the western continental shelf–open ocean boundary of each ocean basin. Prominent WBCs include the Northern Hemisphere Gulf Stream and Kuroshio and Southern Hemisphere Agulhas Current, Brazil Current, and East Australian Current (Talley et al. 2011). Abutting the coastal and shelf ocean, WBCs strongly control the cross-shelf exchange processes for example coastal upwelling and downwelling, and submeso- and mesoscale eddy dynamics. WBC dynamics influence the productivity of the coastal and marine industries, and ecosystem function and health through the exchange of heat, salt, and nutrients. Understanding the controlling dynamics of the boundary currents and their influence on weather to seasonal and longer-term climate variability, regional seas, and marine ecosystem continues to be a key climate research topic (Todd et al. 2019). Knowledge of the influence of WBCs on the coastal and shelf environments is important to marine managers and industries, as WBC variability can drive modifications in productivity, as well as influence the abundance and concentrations of numerous marine species, including those that are commercially exploited or threatened. Reviews of the ocean boundary current systems have repeatedly identified the need for sustained, interdisciplinary observations to meet

 Denotes content that is immediately available upon publication as open access.

Corresponding author: Bernadette M. Sloyan, bernadette.sloyan@csiro.au

DOI: 10.1175/JTECH-D-21-0183.1

© 2023 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy \(www.ametsoc.org/PUBSReuseLicenses\)](https://www.ametsoc.org/PUBSReuseLicenses).

societal needs (Send et al. 2010; Palacz et al. 2017; Todd et al. 2019).

Gould et al. (2013) and Davis et al. (2019) provide a review of ocean observing technologies over the last 100 years. Ocean current mooring arrays, consisting of a sequence of instruments deployed at various depths on a number of moorings across a WBC, directly observe ocean velocity and properties at subhourly temporal resolution (Davis et al. 2019). Early deployments of subsurface moorings in the 1960s and 1970s were challenging with either complete loss of the mooring or very low data return. It was not until the 1980s that successful deployments of longer than a few months and with reliable data return were achieved. Ocean mooring arrays are now a commonly used tool to observe the subsurface ocean properties and have been deployed in most major WBCs, with total records lengths ranging from a few years to just over a decade (Davis et al. 2019; Todd et al. 2019). To date, ocean moorings are still the most effective way to observe WBCs given their relative short spatial scales and high temporal scales (Todd et al. 2019).

WBC mooring arrays consist of subsurface tall moorings that extend from the seafloor to generally 20–30 m below the surface and are deployed for periods between months and years (Davis et al. 2019). Modern instruments, measuring ocean velocity, temperature, salinity, and more recently biogeochemical variables provide high precision and accurate observations (Gould et al. 2013). These instruments provide either an observation at a given depth or observations over some vertical depth range. The moorings do not generally extend to the surface for a variety of reasons, including minimizing mechanical stress on the mooring components from strong surface currents and surface waves/swell, high marine traffic or marine industry activity, and to avoid potential vandalism. Therefore, many mooring arrays do not provide observations at the ocean surface. In addition, WBC moorings can be pushed below the surface by several hundreds of meters, referred to as “blowdown,” by episodic strong current events and/or tidal currents resulting in significant upper-ocean data gaps. Instrument failure and mooring breakages can also lead to the loss of data at any depth between the shallowest and deepest instrument. Finally, data can also be removed if they fail quality assurance and control procedures. For some analyses of the mooring time series data these nonrandom and random data gaps need to be filled.

Various methods have been employed to fill missing mooring data including extrapolating the observed ocean velocity shear, temperature, and salinity gradients; combining mooring velocity data with geostrophic velocity estimates; vertical, temporal, and horizontal interpolation; combining mooring array covariance estimates with least squares regression and sequential multiple regression models (Johnson and McPhaden 1993; Sprintall et al. 2009; Wang et al. 2015; Johns et al. 2001; Kanzow et al. 2006; McMonigal et al. 2020; Li et al. 2020) and where long-time-series data are available missing data may be filled with climatological mean values (Frajka-Williams et al. 2021). The appropriate choice of the method used to fill data gaps is determined by the characteristics (length of time or vertical depth) of the loss of data, the availability of similar and

coherent data or ancillary data, and the planned analyses of the data. For mooring arrays that consist of a number of moorings, such as those deployed in the ocean currents, a combination of data extrapolation, interpolation, and least squares regression methods is commonly applied (Sprintall et al. 2009; Wang et al. 2015). While these methods have been successfully employed to fill data gaps, with the availability of longer (>3 years) mooring time series data and increasing use of machine learning techniques in ocean sciences it is timely to assess the applicability of machine learning to fill data gaps in mooring time series data.

Machine learning, a subset of artificial intelligence, enables a computer to learn from data without being explicitly programmed (Hsieh and Tang 1998; Molnar et al. 2020; Sonnewald et al. 2021). Machine learning is the intersect of computer science and statistics where algorithms are used to recognize “patterns” in the data. These “patterns” are then used for a multitude of purposes including prediction, data classification, and dimensional space reduction (Fradkov 2020). All machine learning techniques are based on determining relationships, or “learning,” from a training dataset and then applying this learned relationship to identify key characteristics or processes. Machine learning techniques, supervised and unsupervised, are becoming widely used within ocean science (Lobo 2009; Sonnewald et al. 2021, and references therein). Machine learning techniques have been applied to in situ ocean observation, satellite data, and model data to investigate ocean circulation, water masses, ocean mixing and to fill data gaps.

Self-organizing maps (SOM), developed in the 1980s by Kohonen (2001), are a machine learning technique based on unsupervised neural networks (Kohonen 2001, 2013). SOM cluster (arrange) high-dimensional datasets onto a lower-, typically two-, dimensional neural map that preserves the topological structure of the data such that neural maps of similar data are near each other and dissimilar classes are separated in the two-dimensional data space. Each neural class is represented by a reference vector in the initial data space. SOM have been widely applied to ocean and marine environment studies (Lobo 2009). In many of these applications SOM have been used for clustering and classifying data contained in two-dimensional classes or images. Lobo (2009) also illustrates novel uses of SOM for segmentation of the seafloor from multibeam data, control of underwater autonomous vehicles, detection of anomalous behavior of ships for maritime vessel traffic systems, and naval operations. More recently SOM have been used to complete observational in situ time series data (Charantonis et al. 2015; Chapman and Charantonis 2017). Here we extend Charantonis et al. (2015) Iterative Completion SOM (ITCOMPSON) method to fill gaps in the velocity, temperature, and salinity time series data from a mooring array deployed in the East Australian Current. We compare the ITCOMPSON data filling method with the more commonly used least squares regression method to assess its performance and suitability to fill data gaps within mooring time series data.

In section 2 we provided a brief description of the East Australia Current (EAC) and detailed information on the

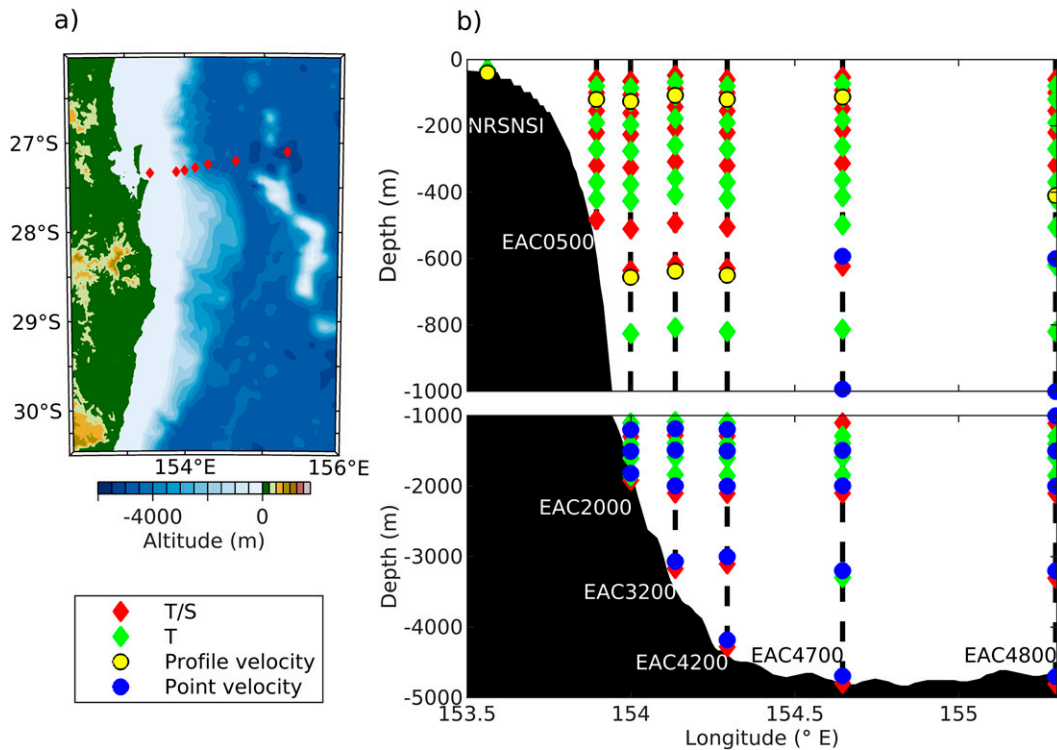


FIG. 1. The EAC mooring array (a) location and (b) vertical distribution of instruments as deployed on the April 2018–September 2021 moorings.

EAC mooring array. We also provide information on the satellite data and other information that is combined with the mooring data to fill the missing data. In section 3 we introduce the ITCOMPSOM and methodology applied to fill data gaps. Validation of the SOM-derived filled velocity, temperature, and salinity mooring data and comparison with the least squares regression method is provided in section 4. A discussion of the ITCOMPSOM method and guide for application to other ocean in situ time series is provided in section 5 and conclusions are presented in section 6.

2. East Australian Current observations

The EAC is the complex, highly energetic western boundary current of the South Pacific Gyre that flows along the east coast of Australia. Due to its broad geographic reach and close proximity to the coast the EAC affects the climate and marine environment from 15° to 42°S. As the strongest ocean current in the region, the EAC and its associated eddy field dominate the marine climate of the east Australian continental shelf and Tasman Sea moving vast amounts of water south from the tropics to the temperate latitudes (Oke et al. 2019). Variability in the EAC's transport of heat, salt, and nutrients modulates the weather systems (i.e., east coast lows; Pepler et al. 2016) that have an acute impact on the heavily populated and industrialized regions of eastern Australia, and composition and functioning of marine ecosystems along the entire east coast (Suthers et al. 2011).

Between 15° and 30°S the EAC is a coherent jet that meanders in an east–west direction onto and off the continental slope, and south of 30°S the EAC is predominantly a series of eddies (Sloyan et al. 2016; Oke et al. 2019). The dynamics and variability of the EAC jet has a strong influence on the characteristics and evolution of the downstream eddy field (Sloyan and O’Kane 2015; Kerry et al. 2018; Kerry and Roughan 2020; Li et al. 2021; Gwyther et al. 2022). In 2012, given the important dynamical constraint that the EAC jet places on the downstream circulation and properties of the coast and the Tasman Sea, the Integrated Marine Observing System (IMOS) and CSIRO Oceans and Atmosphere established a comprehensive deep water mooring array in the EAC jet at approximately 27°S (Fig. 1). Due to the fact that, at 27°S, the EAC is north of the high eddy variability and its flow is relatively uniform and coherent, it is thus suitable for observing using a “picket fence” mooring array. The EAC mooring array is one component of a distributed EAC observing system and part of an international effort to build sustained WBC observations (Todd et al. 2019; Sloyan et al. 2020; www.imos.org, www.ocean-ops.org and www.oceansites.org). The EAC mooring array provides full-depth velocity, temperature, and salinity observations of EAC.

a. Mooring data

The mooring data used in this study consist of velocity, temperature, and salinity observations from the EAC mooring array and velocity and temperature data from a national

TABLE 1. Mooring deployments and years used in this study. Note there is a 22 month break in the mooring data between August 2013 and May 2015. For April 2012–August 2013 (highlighted with a boldface checkmark), we use IMOS SEQ0400 mooring located at $-27^{\circ}19.8'N$ $153^{\circ}52.8'E$ in a water depth of 405 m. The SEQ0400 and EAC0500 are separated by 2.47 km and have a depth difference of 140 m. The EAC3200 mooring has been deployed since 2015.

	Mooring ID						
	NRSNSI	EAC0500	EAC2000	EAC3200	EAC4200	EAC4700	EAC4800
Water depth (m)	63	545	1905	3185	4267	4779	4791
Latitude ($^{\circ}N$)	$-27^{\circ}20.5'$	$-27^{\circ}19.6'$	$-27^{\circ}18.9'$	$-27^{\circ}17.1'$	$-27^{\circ}15.0'$	$-27^{\circ}12.5'$	$-27^{\circ}06.3'$
Longitude ($^{\circ}E$)	$153^{\circ}33.7'$	$153^{\circ}54.0'$	$153^{\circ}59.5'$	$154^{\circ}8.2'$	$154^{\circ}17.9'$	$154^{\circ}38.9'$	$155^{\circ}18.3'$
	Years of data availability						
April 2012–August 2013	✓	✓	✓		✓	✓	✓
May 2015–October 2016	✓	✓	✓	✓	✓	✓	✓
October 2016–April 2018	✓	✓	✓	✓	✓	✓	✓
April 2018–September 2019	✓	✓	✓	✓	✓	✓	✓
September 2019–May 2021	✓	✓	✓	✓	✓	✓	✓

reference mooring station on the continental shelf in-line with the EAC mooring array at North Stradbroke Island (NRSNSI) (Fig. 1, Table 1). Each mooring is instrumented with a combination of acoustic Doppler current profiling (ADCP) instruments of various frequencies (RDI ADCPs 75, 150, and 300 kHz), and point source velocity instruments (Nortek Aquadopp). Over the shelf and continental slope, the ADCPs were deployed on a bottom-deployed tripod (NRSNSI) or in a collocated upward- and downward-looking (at nominally 120 m water depth), and downward-looking (at nominally 620 m water depth) configuration (EAC-0500, 2000, 3200, 4200) providing either full-depth or upper-1000-m velocity observations at vertical resolutions of 4, 8, or 16 m. On the EAC4700 and EAC4800 moorings, ADCPs provide velocity observations over the upper 600 and 500 m of the water column, respectively. Point source velocity data were obtained below 1000 m to the seafloor at vertical resolutions of between 500 and 1000 m, except for EAC4700 and EAC4800 where point-source velocity is obtained below 600 m. Temperature observations were obtained from Sea-Bird Electronic (SBE) SBE39-plus, SBE37-SMP MicroCAT, and Starmon Mini instruments. Salinity observations were collected with Sea-Bird Electronic SBE 37-SMP MicroCAT. Temperature and salinity observations are at 20 m vertical resolution from approximately 20 to 200 m, at 50 m vertical resolution from 200 to 500 m, at 100–200 m vertical resolution from 500 to 1000 m, and at 200–300 m vertical resolution from 800 to 2000 m. However, salinity observation vertical resolution varies from 40 to 200 m in the upper 2000 m for the 2012/13, 2015/16, and 2016–18 mooring array deployments. Below 2000 m to the seafloor, temperature and salinity are observed at a vertical resolution of 500–1000 m. Temperature and salinity vertical resolutions are similar for all moorings. The vertical resolution of the instruments are chosen to best observe the complex vertical structure of the velocity, temperature, and salinity including a subsurface velocity maximum that is generally found between 60 and 120 m, the strong seasonal thermocline and halocline, the permanent thermocline, and the subsurface salinity maximum and minimum.

The first deployment of the EAC mooring array in 2012/13 consisted of seven moorings that extended from the continental shelf in 200 m of water to the abyssal ocean in a water depth of 4797 m (Sloyan et al. 2016). In 2015 a redesigned EAC mooring array consisting of six moorings was deployed with moorings in 500, 2000, 3200, and 4200 m water depth on the continental shelf and slope, and moorings in water depths of 4700 and 4800 m in the adjoining abyssal plain. Here we combine the 2012/13 and 2015–21 EAC mooring arrays to build a consistent time series of mooring data from 2012 to 2021, excluding the data gap between 2013 and 2015 when the mooring array was not deployed (Table 1).

Given the EAC mooring array design changes between deployments 2012/13 and 2015 to the present, the following decisions were made: We exclude the 2012/13 moorings at 200 m (SEQ200) and 1500 m as these mooring were not continued when the array was reestablished in 2015. Since 2015 the EAC mooring array has maintained a mooring in 500 m (EAC0500), rather than 400 m (SEQ0400) as was deployed in 2012/13. The distance separating these mooring locations is only 2.47 km. We combine the 2012/13 SEQ0400 deployment with EAC0500 to provide an upper continental slope observational record across the 2012–21 time period. To extend the EAC mooring array onto the continental shelf we include the IMOS NRSNSI mooring (water depth = 63 m). The NRSNSI site has been continuously occupied during the period of EAC mooring array.

The quality assured and controlled FV01 IMOS EAC instrument data (Cowley 2021; Lovell and Cowley 2022a,b; Cowley 2022a,b) and SEQ0400 and NRS North Stradbroke Island data were downloaded from the IMOS Australian Ocean Data Network and compiled by mooring. The compiled hourly FV01 data are available from <https://doi.org/10.25919/xkgx-zy14> (Sloyan and Cowley 2022). Surface reflection errors result in ADCP's providing velocity to within two bins of the ocean surface; thus, our shallowest ocean velocity data are at 20 m below the surface. Where data are available, the data were interpolated onto a standard depth grid between 0 and 5000 m with a vertical interval of 10 m in the upper 400

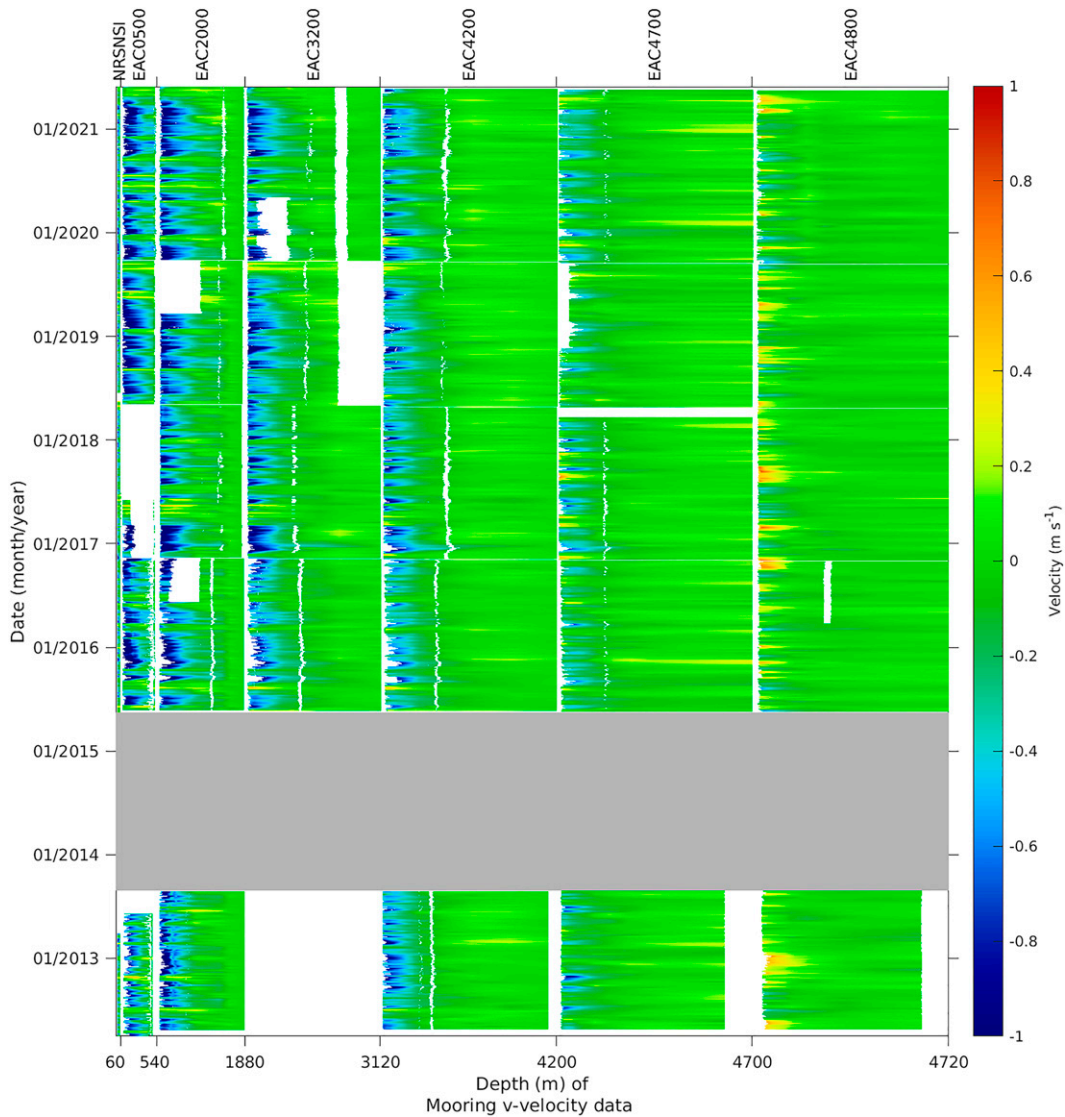


FIG. 2. Data matrix of the aggregated daily v -velocity-component (m s^{-1}) vertical profiles for the seven moorings of the EAC array. The upper x -axis label identifies each mooring and the lower x axis shows the maximum depth of each moorings vertical profile. That is, vertical profiles are NRSNSI: 0–60 m; EAC0500: 0–540 m; EAC200: 0–1880 m; EAC3200: 0–3120 m; EAC4200: 0–4200 m; EAC4700: 0–4700 m; and EAC4800: 0–4720 m. Note that 0 m is not labeled but follows directly after preceding mooring and starts the array for NRSNSI. Negative v velocity is southward. Missing data are shown as white gaps in the time series. The 2013–15 period when the mooring array was not deployed is identified by gray.

and 20 m from 400 m to the seafloor. A 5-day filter was applied to the mooring data to remove tides and other high-frequency processes, and all data were then interpolated to a common daily time stamp. The unfilled depth and time gridded data are available from Sloyan et al. (2021, <https://doi.org/10.25919/a8j3-zh92>). The individual gridded mooring data form a matrix of daily observations across the EAC from the coastal mooring (NRSNSI) to offshore (EAC4800) (Figs. 2–4).

b. Satellite data

We use satellite altimeter sea level anomaly (SLA) and satellite derived temperature and salinity data in this study. The

satellite SLA is used as additional data and appended to the daily mooring velocity matrix. The sea surface temperature (SST) and salinity (SSS) data provide the daily surface boundary condition for the mooring temperature and salinity data.

IMOS SLA are obtained from the Australian Ocean Data Network (portal.aodn.org.au). This product uses coastal tide gauge data interpolated around the Australian coastline to extend the satellite SLA across the continental shelf (Deng et al. 2011). This extension to the coast is particularly important at 27°S as the EAC abuts a narrow continental shelf. The daily $0.2^\circ \times 0.2^\circ$ gridded data are interpolated to mooring location and time stamp.

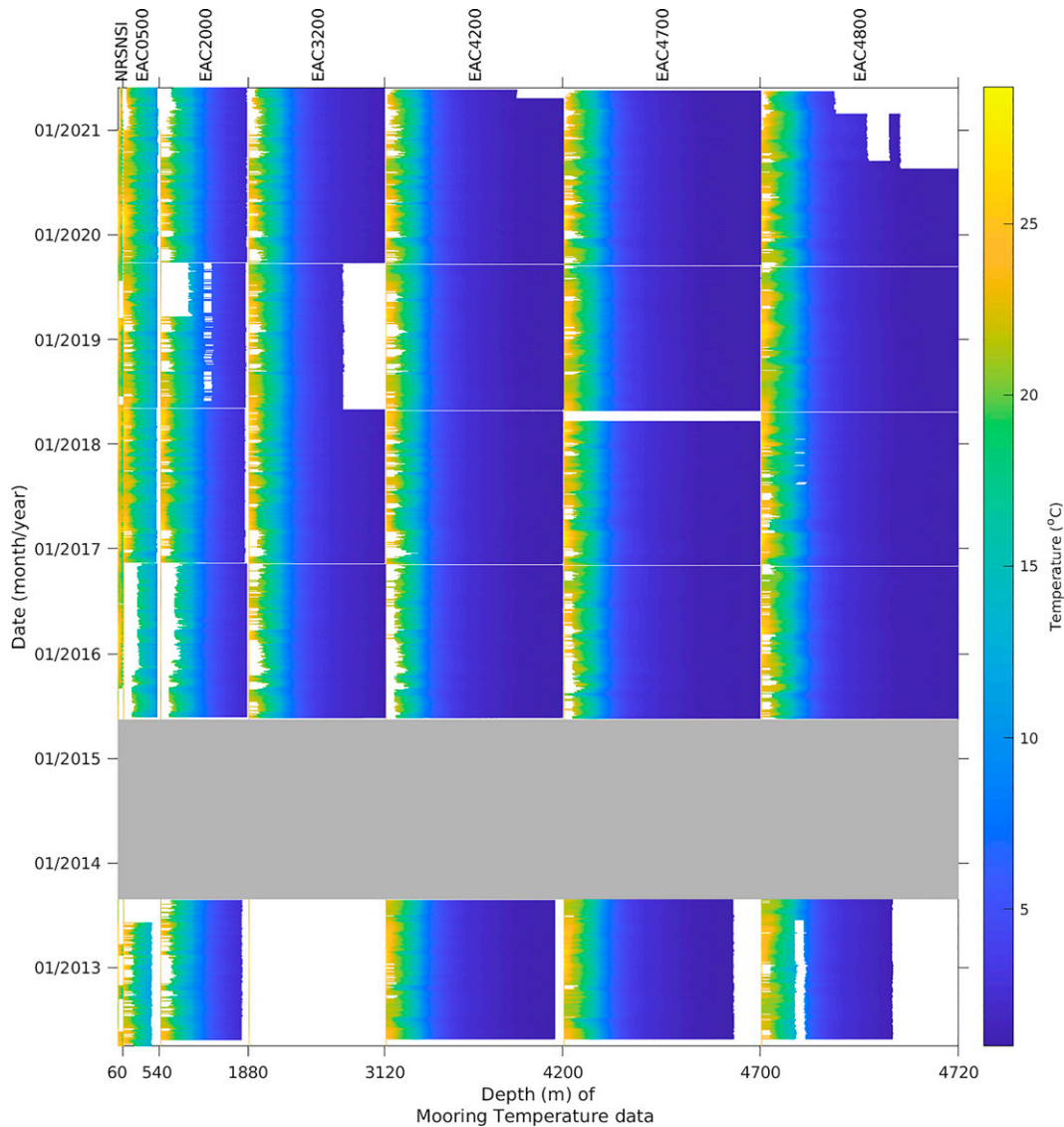


FIG. 3. As in Fig. 2, but for temperature ($^{\circ}\text{C}$) data.

The NOAA/NESDIS/NCEI Daily Optimum Interpolation Sea Surface Temperature (OISST), version 2.1 data (Huang et al. 2021) and OCEAN/IPSL SSS L3 version 5 data from Soil Moisture Ocean Salinity (SMOS) satellite (Boutin et al. 2018, 2020) are linearly interpolated to the mooring location and daily time stamp. The SST and SSS data are added as a surface boundary condition to the mooring subsurface temperature and salinity data.

c. EAC coherence and seasonality

The NRSNSI and EAC moorings are not evenly distributed with distance along the mooring line. The NRSNSI mooring is located 30 km inshore of the EAC0500 mooring. The EAC moorings horizontal separation is between 10 and 16 km down the continental slope (EAC0500 to EAC4200) and then increases to 35 and 60 km across the abyssal plain between EAC4200, EAC4700, and EAC4800, respectively. The relatively

short separation of the moorings on the continental slope resolves the complex cross-slope dynamics (temporal and spatial) of the jet core and the increased spacing of the mooring over the abyssal plain resolves the offshore recirculation and periods when the EAC completely detaches from the continental slope and meanders eastward. To represent the coherence of the EAC-jet across the mooring line and the varying separation of the moorings between NRSNSI and EAC4800 we define a distance (Dist) variable that is the distance of the moorings from a geographical point slightly to the west of the NRSNSI mooring location.

Observational- and model-based studies suggest that the EAC has a volume and upper-ocean temperature seasonal cycle (Ridgway and Godfrey 1997; Wood et al. 2016; Kerry and Roughan 2020). These studies have shown that southward EAC volume transport has a maximum from December to February and a minimum between May and September. To

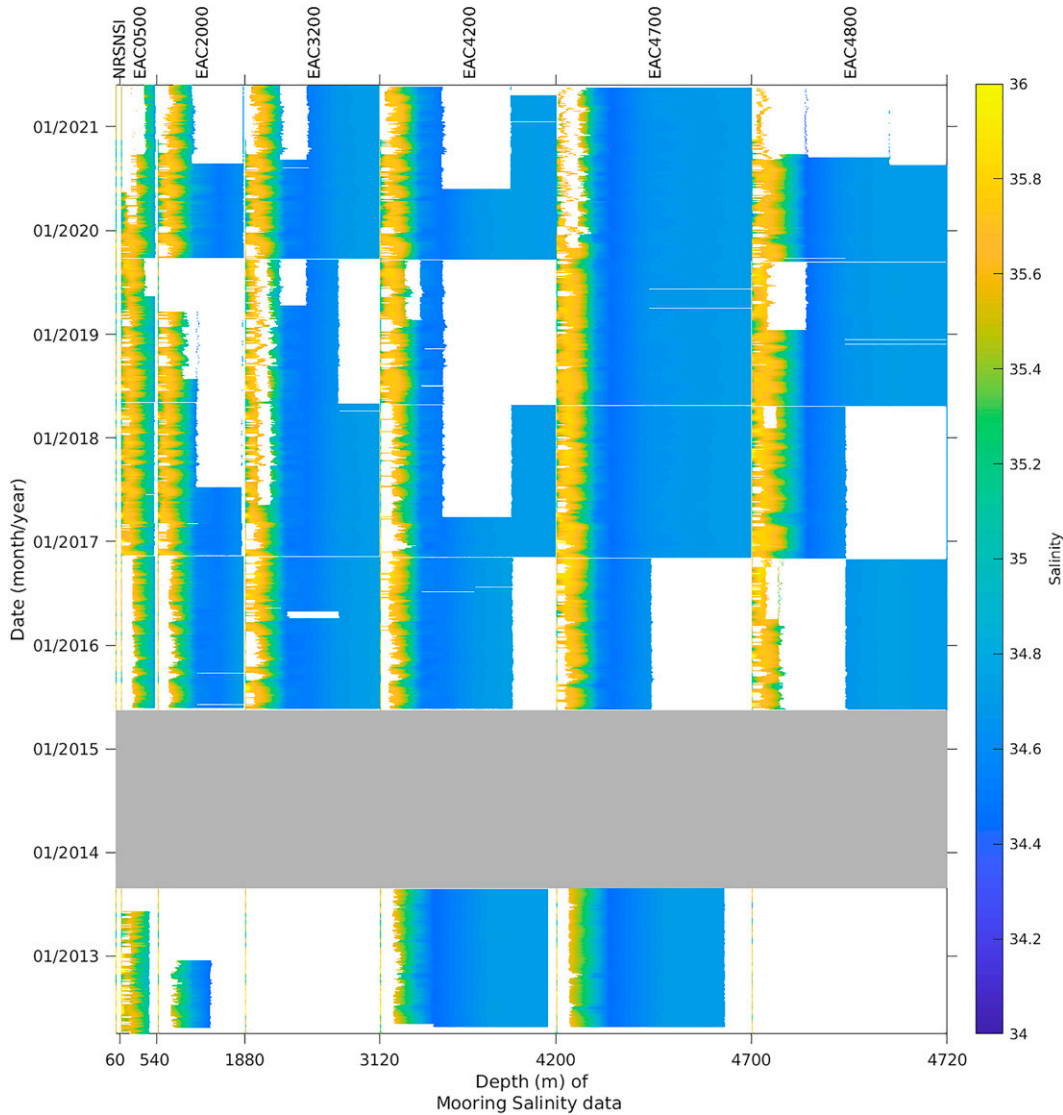


FIG. 4. As in Fig. 2, but for salinity data.

represent this seasonality, we define a day of year (DoY) variable.

LEAST SQUARES REGRESSION

Following Sprintall et al. (2009) and Wang et al. (2015), a least squares regression model is developed to fill mooring data gaps in the velocity time series. As with previous studies, we first fill small depth (<40 m) and short time (<10 days) data gaps using linear interpolation. The linear regression fit, $\mathbf{U}_{\text{pred}} = a\mathbf{U}_{\text{NRSNSI}} + b\mathbf{U}_{\text{EAC0500}} + c\mathbf{U}_{\text{EAC2000}} + d\mathbf{U}_{\text{EAC3200}} + e\mathbf{U}_{\text{EAC4200}} + f\mathbf{U}_{\text{EAC4700}} + g\mathbf{U}_{\text{EAC4800}}$, where \mathbf{U}_{pred} is the predicted zonal (u) or meridional (v) velocity at a mooring and $\mathbf{U}_{\text{NRSNSI}, \text{EAC0500}, \text{EAC2000}, \text{EAC3200}, \text{EAC4200}, \text{EAC4700}, \text{EAC4800}}$ is the zonal (u) or meridional (v) velocity at surrounding mooring sites and coefficients $a, b, c, d, e, f,$ and g are determined by linear damped least squares regression from training (predictor)

data that are taken from the observed velocity data. That is, predicted mooring velocity data, at each depth, is expressed as a linear combination of the normalized velocity data whose coefficients are determined using a least squares fit. The predicted velocity is then used to fill data gaps in a moorings velocity time series.

3. Self-organizing maps

SOM, developed in the 1980s by Kohonen (2001), is a machine learning technique based on unsupervised neural networks. SOM are used to cluster high-dimensional datasets arranging (or organizing) them on a lower-, typically two-, dimensional neural map that preserves the topological structure of the data such that neural classes of similar data are near each other and dissimilar classes are separated in the two-dimensional data space. Each neural class is represented by

a referent vector in the initial data space. For a thorough introduction to machine learning techniques and their application to oceanography we suggest investigation of [Hsieh and Tang \(1998\)](#), [Lobo \(2009\)](#), [Kohonen \(2001\)](#), [Molnar et al. \(2020\)](#), and [Sonnewald et al. \(2021\)](#). In particular, [Sonnewald et al. \(2021\)](#) provide a brief introduction to machine learning techniques and reference suggestions for further reading.

Following [Puissant et al. \(2021\)](#), [Chapman and Charantonis \(2017\)](#), and [Charantonis et al. \(2015\)](#) we apply an ITCOMPSOM method to fill data gaps in the mooring time series. The missing data are randomly distributed throughout the time period and there are no instances, except for the period between 2013 and 2015, when there are no velocity, temperature, or salinity observations of the EAC ([Figs. 2–4](#)). Combining the velocity, salinity, and temperature data from each individual mooring enables us to provide daily EAC observations for 2012–21 to train the ITCOMPSOM and fill the mooring data gaps.

a. Data matrices

For the 2012–21 period the percentage of missing mooring data varies from between 7% and 46% for velocity, 6%–30% for temperature, and 20%–54% for salinity, excluding the NRSNSI mooring where we do not use the subsurface salinity observations ([Table 2](#)). The largest percentage of missing data is found at the moorings located on the shelf and continental slope where the moorings experience the greatest mechanical stress due to them being located in the mean EAC jet position and most marine industry pressure due to their proximity to the coast.

Following [Vesanto et al. \(2000\)](#) SOM toolbox documentation, the daily (t) velocity [$\mathbf{U}(u, v)$] and temperature (T) and salinity (S) data for each mooring and ancillary data are compiled into data matrices (**DAT_U** and **DAT_TS**), spanning the period April 2012–May 2021, respectively. For the velocity data matrix ancillary data include SLA, jet coherence (Dist), and seasonality (DoY), and for the temperature and salinity data matrix only jet coherence and seasonality are included. The data matrices then have the form

• Velocity

$$\mathbf{DAT_U} = \begin{bmatrix} \mathbf{U}_{\text{moorings}_1} & \text{SLA}_1 & \text{Dist} & \text{DoY}_1 \\ \mathbf{U}_{\text{moorings}_2} & \text{SLA}_2 & \text{Dist} & \text{DoY}_2 \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{U}_{\text{moorings}_t} & \text{SLA}_t & \text{Dist} & \text{DoY}_t \end{bmatrix},$$

where t is sequential daily time from 2012 and $\mathbf{U} = (u, v)$. $\mathbf{U}_{\text{moorings}_t} = [\mathbf{U}_{\text{NRSNSI}_t}, \mathbf{U}_{\text{EAC0500}_t}, \mathbf{U}_{\text{EAC2000}_t}, \mathbf{U}_{\text{EAC3200}_t}, \mathbf{U}_{\text{EAC4200}_t}, \mathbf{U}_{\text{EAC4700}_t}, \mathbf{U}_{\text{EAC4800}_t}]$ is a matrix of the aggregated daily vertical velocity profiles of each mooring, as shown in [Fig. 2](#) for the v -velocity component.

TABLE 2. Percentage (%) of missing velocity, temperature, and salinity data for each mooring for the combined time period 2012/13 and 2015–21. For EAC3200, the percentage of missing data between 2015 and 2021 when the mooring was deployed is shown in parentheses. Note also that for NRSNSI the only salinity data available are the SSS and the subsurface salinity data are not used in this study.

	Velocity (%)	Temperature (%)	Salinity (%)
NRSNSI	46	30	86
EAC0500	39	25	33
EAC2000	14	15	48
EAC3200	33 (20)	28 (14)	34 (24)
EAC4200	7	6	31
EAC4700	8	7	20
EAC4800	8	14	54

• Temperature and salinity

$$\mathbf{DAT_TS} = \begin{bmatrix} T_{\text{moorings}_1} & S_{\text{moorings}_1} & \text{Dist} & \text{DoY}_1 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ T_{\text{moorings}_t} & S_{\text{moorings}_t} & \text{Dist} & \text{DoY}_t \end{bmatrix},$$

where $T_{\text{moorings}_t} = [T_{\text{NRSNSI}_t}, T_{\text{EAC0500}_t}, T_{\text{EAC2000}_t}, T_{\text{EAC3200}_t}, T_{\text{EAC4200}_t}, T_{\text{EAC4700}_t}, T_{\text{EAC4800}_t}]$ and $S_{\text{moorings}_t} = [S_{\text{NRSNSI}_t}, S_{\text{EAC0500}_t}, S_{\text{EAC2000}_t}, S_{\text{EAC3200}_t}, S_{\text{EAC4200}_t}, S_{\text{EAC4700}_t}, S_{\text{EAC4800}_t}]$ are the aggregated daily vertical temperature and salinity profiles of each mooring, as shown in [Figs. 3 and 4](#). We remind the reader that satellite SST and SSS are added to the mooring data to provide surface temperature and salinity observations.

b. Determining the SOM velocity, salinity, and temperature classes

The ITCOMPSOM method, combining approaches previously used in [Puissant et al. \(2021\)](#) and [Charantonis et al. \(2015\)](#), determines the SOM classes that are then used to fill the missing mooring velocity and temperature and salinity data. The ITCOMPSOM iterative completion of the missing data has been shown to be better than basic SOM methods for datasets that contain nonrandom missing data ([Puissant et al. 2021](#)). ITCOMPSOM inputs missing values of a data vector several times from progressively larger topological maps that combine previously completed data with new data (with missing values) at each iteration. We used the [Vesanto et al. \(2000\)](#) SOM toolbox for MATLAB 5 in this study.

The classes are determined by training the SOM on the available mooring data at each daily time step. Before using the **DAT_U** and **DAT_TS** matrices in the SOM training phase we manipulate the data matrices by first removing the time period between September 2013 and April 2015 when the EAC mooring array was suspended, and then sort the daily data vectors in descending order of observed mooring data (i.e., the first row of the matrix has the least mooring data gaps and the last row has the most missing data). The

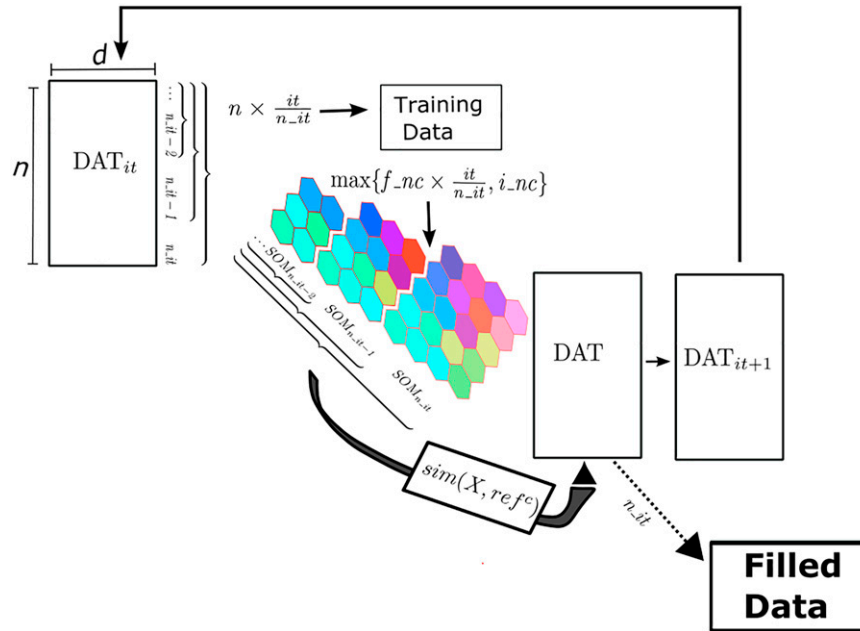


FIG. 5. Schematic of the ITCOMPSOM methodology applied to fill gaps in the mooring velocity, salinity, and temperature data. **DAT** is the data matrix, n is the number of daily time vectors, and d is observational variables, “it” is the iteration number, n_it total number of iterations, i_nc is the initial number of SOM classes, f_nc final number of classes, and $sim(\mathbf{X}, ref_c)$ is the similarity function for best fit class for missing data in \mathbf{X} .

resultant data matrix (**DAT**) dimension is $\mathbf{n} \times d$, where \mathbf{n} is the reordered daily time vector and d is the mooring and ancillary observations. Finally, we normalize each variable by its variance. We now only refer to **DAT**, which is a matrix of the reordered and normalized daily velocity or temperature and salinity data, as we describe the ITCOMPSOM method used to determine the SOM.

We build the SOM used to fill the missing data by iterating **DAT** over n_it iterations while progressively increasing the number of SOM classes from an initial number of classes i_nc to a final number of classes f_nc . We choose $n_it = 25$ which results in an incremental increase of 85 days ($n \times it/n_it$) for each iteration. The initial value of SOM classes, $i_nc = 200$ for velocity and 100 for salinity and temperature, were selected using the Davies–Bouldin index (Davies and Bouldin 1979). The maximum number of SOM classes at the final iteration is $f_nc = 1300$ for velocity and 550 for salinity and temperature.

At each iteration, **DAT** contains a larger portion of \mathbf{n} ($n \times it/n_it$) and increasing number of SOM classes nc_it , from i_nc to f_nc and training successively larger data classes that combine previously completed data and new data with missing values. That is, at each iteration step we consider a matrix **DAT_it** and estimate missing data in **DAT** from the SOM nc_it . The filled data matrix **DAT_it** together with additional unfilled data from **DAT** are used in the next iteration, **DAT_it + 1** (Fig. 5). The iteration loop is continued until it = n_it and $nc = f_nc$ and the final SOM classes are used to fill all missing mooring data in **DAT**. Therefore, missing data are filled several times during the iterative process until the entire dataset is filled at the last iteration n_it . Finally, the completed

DAT matrix is denormalized and reordered to sequential time and velocity, temperature, and salinity data for each mooring are extracted. The 30 m velocity value is taken as a constant and used to extend the velocity profile to the ocean surface.

The completion of **DAT_it** is achieved by assigning each vector \mathbf{X} of **DAT** to a particular SOM class, c . Following Charantonis et al. (2015) and Chapman and Charantonis (2017) we define a similarity function, $sim(\mathbf{X}, ref_c)$ to choose the most appropriate SOM class, known as the best matching unit (BMU). $sim(\mathbf{X}, ref_c)$ weights the Euclidean (data) distance by the correlation between the missing data and the available data such that

$$sim(\mathbf{X}, ref_c) = \sum_{j \in obs_i} \left[1 + \sum_{j \in missing_i} (\mathbf{cor}_{ij}^c)^2 \right] \times (x_i - ref_i^c)^2, \tag{1}$$

where each SOM class at iteration it is represented by c , x_i is the missing data in \mathbf{X} , ref_c is the mean of all training data in SOM class c , and \mathbf{cor}_{ij}^c is the correlation matrix between the missing and the mean of all the observed training data within c . Including the correlation between missing and observed data, we chose the SOM class that has the highest correlation of the observed and predicted velocity, temperature, and salinity vertical profile from a choice that simply considers the Euclidean distance.

4. Filled velocity, temperature, and salinity mooring data

a. Validation of ITCOMPSOM

We validate the ITCOMPSOM method to fill missing velocity, temperature, and salinity data in two ways: randomly

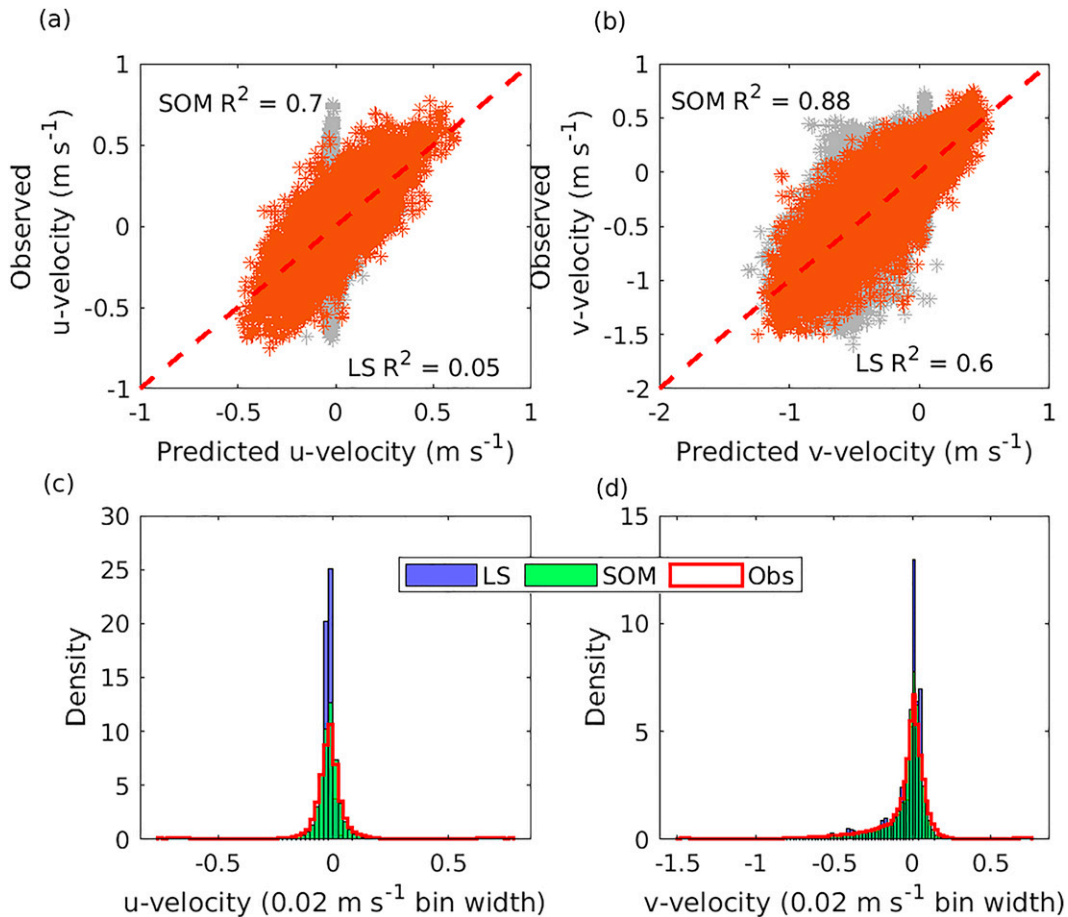


FIG. 6. Predicted ITCOMPSOM (orange) and least squares regression (gray) vs observed velocity (m s^{-1}) (a) zonal (u) and (b) meridional (v) for 20% randomly selected data. The density distribution of the (c) zonal (u) and (d) meridional (v) velocity observed (red), ITCOMPSOM (SOM, green filled) and least squares regression (LS, purple filled) predicted data. Twenty percent of the observed data was withheld. The line of best fit (red dashed line) is shown for reference. The ITCOMPSOM and least squares R^2 values for the u - and v -velocity components are provided in (a) and (b).

withholding parts of fully known profiles and withholding large continuous time periods of known profiles. ITCOMPSOM reconstructed data are then compared to the known withheld data. These two validation procedures mimic the sources of data loss that result in data gaps in the mooring time series. Data that fail the quality assurance and control processes result in random data loss in the time series and failure or loss of an instrument results in large vertical and temporal blocks of missing data. We assess the performance of the ITCOMPSOM against the more commonly used least squares regression data filling method, to which the same validation procedures have been applied.

1) VELOCITY

To assess the ability of the ITCOMPSOM to reconstruct randomly missing data we withhold 20% of known data from the data matrix and compared the known “true” data with the ITCOMPSOM reconstructed data.

This 80:20 split ratio of training to validation samples has become convention in machine learning literature, and although investigation by numerous studies have attempted to find optimal split ratios no firm conclusions have been reached for practical application outside of procedures with where few parameters (order 10) are to be fit (Joseph 2022). As the ITCOMPSOM method fits on the order of thousands of parameters (classes), we have followed convention and settled on 80:20 split, which is both simple to implement and allows a straightforward comparison between ITCOMPSOM method and linear least squares regression.

We perform the same validation procedure to the least squares regression method to compare with the ITCOMPSOM results. The coefficient of determination R^2 for the ITCOMPSOM and least squares regression for the u -velocity components are 0.70 and 0.053, respectively, and the root-mean-square error (RMSE) of the true and predicted data is 0.038 and 0.68 m s^{-1} , respectively (Fig. 6a). The ITCOMPSOM captures the density distribution of the withheld u -velocity data much better than the least squares

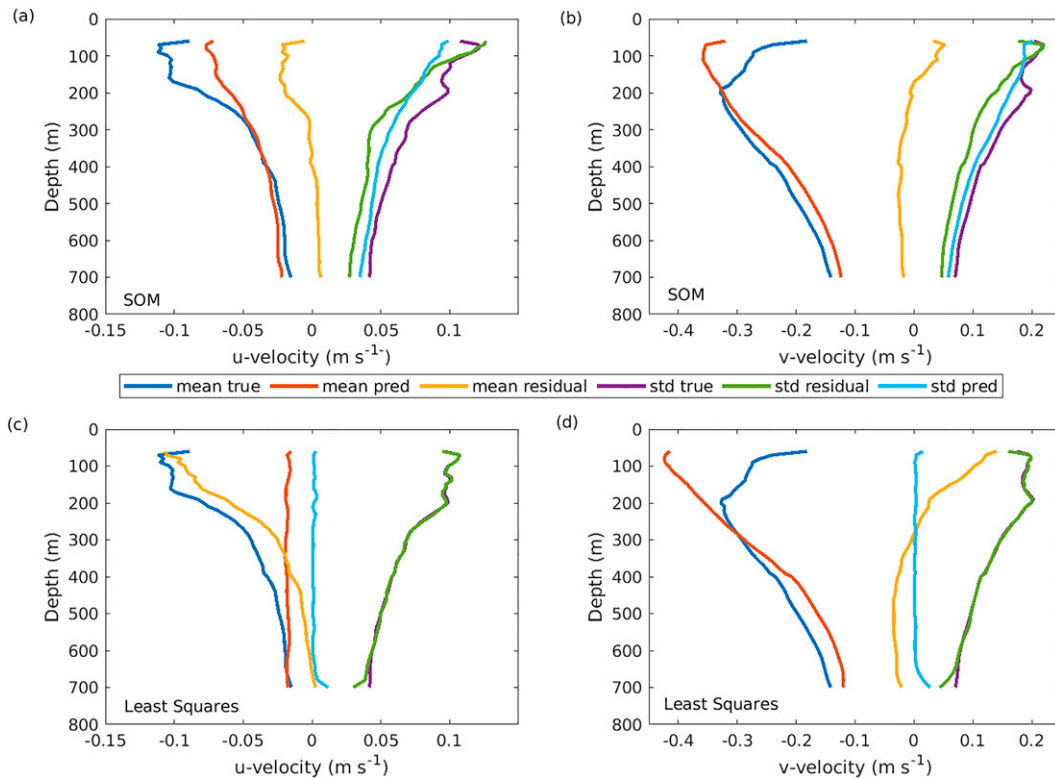


FIG. 7. Comparison of the observed mean u - and v -velocity profile for the (a),(b) ITCOMPSOM and (c),(d) least squares reconstruction of the 100 days of withheld data between 60 and 700 m of the EAC4200 mooring. The u - and v -velocity mean true profile (blue line), reconstructed profile (red line), and the residual (true minus reconstruction) profile (yellow) are shown for the ITCOMPSOM and least squares method. Also shown are the standard deviation for the true (purple line), reconstructed (cyan line), and residual (green line) velocity profiles.

regression method (Fig. 6c). The v -velocity component R^2 values of the ITCOMPSOM and least squares method are 0.88 and 0.60, respectively, and both methods capture well the density distribution of the withheld data (Figs. 6b,d). The v -velocity RMSE error of the ITCOMPSOM and least squares method are 0.064 and 0.12 m s^{-1} , respectively. The RMSE of the u and v velocity of the true and ITCOMPSOM reconstructed data is approximately half that achieved by the least squares method.

We also compare the true speed (velocity magnitude) and direction with the ITCOMPSOM and least squares reconstructions for speeds of greater than 0.05 m s^{-1} . The R^2 coefficient of the ITCOMPSOM reconstruction is 0.85 and for the least squares reconstruction is 0.60 for current speed and the mean angle difference is 1.4° and 1.8°, respectively. The RMSE of speed greater than 0.05 m s^{-1} for the ITCOMPSOM and least squares method is 0.07 and 0.12 m s^{-1} , respectively. The RMSE of true and reconstructed data for the least squares method is nearly double that achieved by the ITCOMPSOM method.

To assess whether or not the ITCOMPSOM is overfitting the data we randomly select 20% of the known data used in the training dataset and compare this to the ITCOMPSOM predicted data. The RMSE is 0.033 and 0.057 m s^{-1} for the u - and v -velocity components, respectively. These RMSE values are in agreement with the RMSE of the validation data suggesting that the ITCOMPSOM is not overfitting the data.

To observe the upper-ocean velocity (30–1000 m) the mooring array uses profiling velocity instruments (Fig. 1). These instruments collect ocean velocity data over depth ranges of 150 m (RDI 300 kHz), 300 m (RDI 150 kHz), and 500 m (RDI 75 kHz). (In 2016 to minimize blowdown data loss due to strong currents upward-looking RDI 150 kHz ADCPs were deployed at a nominal depth of 120 m on the EAC2000, EAC3200, and EAC4200 moorings replacing previously used RDI 300 kHz ADCP instruments. The increased profiling range of the RDI 150 kHz ADCP significantly reduced the loss of data during strong upper-ocean current events.) The use of profiling velocity instruments has greatly improved the vertical resolution of ocean velocity data; however, if they fail no velocity data are collected for the entire depth range and time period until the mooring is serviced—see Fig. 2 EAC0500 and EAC2000 data record. To assess the ITCOMPSOM’s and least squares method’s ability to fill velocity data loss due to the failure of a profile velocity instrument we withheld known velocity data between 60 and 700 m for 100 days from the EAC4200 mooring. The ITCOMPSOM and least squares reconstructions are then compared to the true velocity profile data.

The ITCOMPSOM reconstructed mean u - and v -velocity profiles reproduce the vertical structure of the observed velocity profile very well below 200 m (Figs. 7a,b). Above 200 m

the ITCOMPSOM, while still capturing the vertical structure, under- or overestimate the u - and v -velocity strength when compared to the true observations (Figs. 7a,b). The least squares method reconstruction also captures well the structure of the v velocity below 300 m of the true profile, and similar to the ITCOMPSOM overestimates the strength to the upper-ocean velocity (Fig. 7d). However, the least squares method reconstruction of the u -velocity profile is poor (Fig. 7c). A comparison between the true and ITCOMPSOM and least squares reconstructed profiles is shown by examining the residual (true minus reconstruction) vertical profile and velocity profile standard deviation (Fig. 7). The ITCOMPSOM and least squares mean u - and v -velocity profile residuals are 0.009 and 0.020 m s^{-1} and 0.039 and 0.041 m s^{-1} , respectively. The ITCOMPSOM residual profile is smaller than the least squares residual profiles at all depth showing that the ITCOMPSOM provides a better reconstruction of the data. As mentioned, the EAC is a dynamic current with significant temporal variability, the profile of velocity standard deviation provides a measure of this variability. The ITCOMPSOM reconstruction has a similar standard deviation to the true profile and the least squares reconstruction does not recreate the true variability of the velocity. This shows that the ITCOMPSOM captures the variability of the current much better than the least squares method.

The difference in the ability of both the ITCOMPSOM and least squares method to reproduce the u - and v -velocity components may be explained by the magnitude and the spatial and temporal variability of the velocity across the EAC mooring array. The EAC is a southward flowing current with the mean v -velocity component being more than 2 times larger than the u -velocity component and its variability dominated by the east–west meandering of the current jet (Sloyan et al. 2016). The current orientation and variability explains the higher R^2 coefficient for the v -velocity than the u -velocity component for the validation of randomly withheld known data. Both the ITCOMPSOM and least squares methods predict the v -velocity component more accurately than the u -velocity component; however, the ITCOMPSOM method RMSE error is approximately half that of the least squares method and significantly better reconstructs the density distribution of the u -velocity component. While both the ITCOMPSOM and least squares methods constructed the u velocity less well than the v component, the ITCOMPSOM provides a much better reconstruction of the u velocity when compared to the least squares method. The reduced accuracy of the least squares method to reconstruct the nondominant velocity component was also noted by Wang et al. (2015), and Sprintall et al. (2009) use the least squares method to only reconstruct the dominant along-stream velocity of the Indonesian Throughflow.

2) TEMPERATURE AND SALINITY

As the ITCOMPSOM is found to provide an improved method to fill missing velocity data, when compared to a previously used method, we now assess the ITCOMPSOM ability to fill the missing gaps in the temperature and salinity time series data. We validate the salinity and temperature filled data

by withholding 20% of known data from the training data and compare the true and reconstructed data. The ITCOMPSOM predicted salinity and temperature data have R^2 coefficients of 0.99 and 1, respectively, and an RMSE between true and reconstructed data of 0.04 and 0.38°C, respectively (Fig. 8). The density distribution of the salinity and temperature is complex reflecting the vertical structure of each property. The 20% withheld data adequately samples this complex vertical structure. The ITCOMPSOM density distribution agrees with the complex observed data distribution (Fig. 8). This shows that the ITCOMPSOM reconstruction is able to reconstruct the vertical distribution of the salinity and temperature profiles including the salinity minima and maxima layers and the sharp temperature thermocline.

Finally, to assess whether or not the ITCOMPSOM is overfitting the data we randomly select 20% of the known data used in the training dataset and compare this to the ITCOMPSOM predicted data. The RMSE 0.05 and 0.55°C for salinity and temperature, respectively. These RMSE values are in agreement with the RMSE of the validation data suggesting that the ITCOMPSOM is not overfitting the data.

b. Filled observational time series

Section 4a shows that the ITCOMPSOM is an appropriate machine learning method to fill missing data in a mooring time series data. We now use all available mooring velocity and salinity and temperature data, including those previously withheld in the validation step, to train the SOM. The ITCOMPSOM filled data are a mean of daily profiles that are attributed to the best fitting class determined by the similarity function. We use the best fitting mean profile to fill the missing data, excluding the 2013–15 period when the EAC mooring array was not deployed. For each mooring the observed and filled daily time series data are available from the CSIRO Data Portal (Sloyan et al. 2021; <https://doi.org/10.25919/a8j3-zh92>).

The EAC0500 and EAC2000 mooring missing data are greater than 14% for all variables. These mooring are key to identifying the EAC's inshore edge and jet location on the continental slope. The ITCOMPSOM filled velocity data are consistent with a temporal period of the data gap at the mooring itself and with the adjacent mooring (Fig. 9). Similarly, the salinity and temperature filled data are consistent with the observed data temporally and spatially variability (Fig. 10). Finally, we show that the EAC observational-only and ITCOMPSOM filled mooring time series vertical residual profile for the full-depth profile for EAC0500 and EAC2000 and upper 2000 m for the remaining moorings are small (Figs. 11–13). The velocity residual is shown for depth greater than 30 m, as above 30 m due to surface reflection error no ADCP velocity is available. The property residuals are all small, being generally less than 0.01 m s^{-1} for velocity, 0.1°C for temperature, and 0.02 for salinity. The residuals are much smaller than the property standard deviation of each mooring.

5. Discussion

We have successfully implemented a machine learning method, ITCOMPSOM, to fill missing data in the velocity,

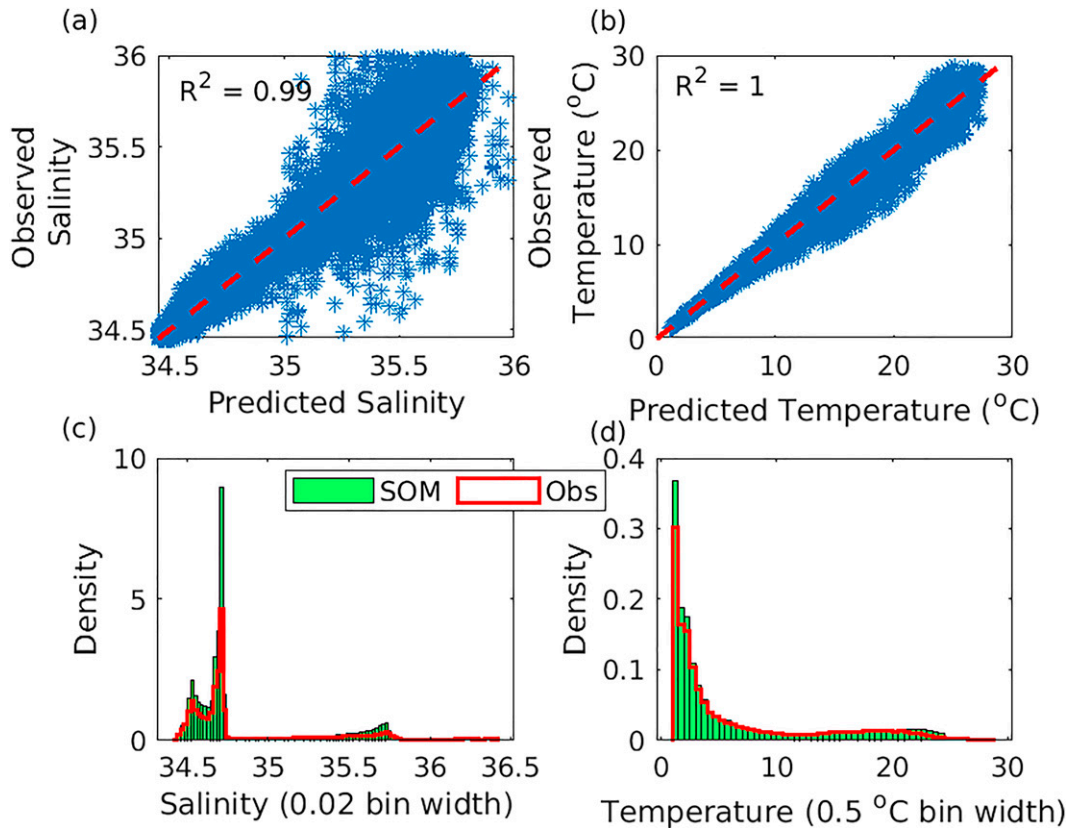


FIG. 8. Predicted vs observed (a) salinity and (b) temperature ($^{\circ}\text{C}$) for the randomly selected data withheld from ITCOMPSOM. The density distribution of the ITCOMPSOM predicted and observed data (c) salinity and (d) temperature. Twenty percent of the observed data was withheld.

temperature, and salinity observational time series of the East Australian Current. We have shown that the ITCOMPSOM outperforms the more commonly used combined extrapolation, interpolation, and least squares regression method. The root-mean-square error of the true and predicted ITCOMPSOM data is approximately half that achieved by the least squares method and the ITCOMPSOM velocity profile standard deviation better represents the observed profile standard deviation when compared to the least squares method. The profile residual of the observational-only and ITCOMPSOM filled data are small. To aid future application of the ITCOMPSOM method for in situ ocean time series data filling, we now discuss how to choose various SOM parameters and benefits and limitations of the method.

In this study we have utilized the often used and well-documented Vesanto et al. (2000) SOM toolbox for MATLAB 5. The toolbox and documentation was downloaded from <http://www.cis.hut.fi/projects/somtoolbox>. We used MATLAB R2020b Update (6 9.9.0.1718557) on a 64 bit Dell PowerEdge R360 with a 2.60 GHz 16-core processor Intel Xeon CPU and 512 GB of memory to run the training and reconstruction code components. Clock time was 0.87 h to complete the velocity data filling calculation, and temperature and salinity calculation are complete in half this time. We were also able to run the code on a MacBook Pro with 2.3 GHz 8-core Intel core i9 processor and 32 GB memory;

however, the clock time during the SOM training phase was significantly longer on the MacBook Pro. The SOM toolbox documentation provides information on the minimum system requirement required to run the software.

Following the SOM toolbox manual, we setup our data matrices ensuring that each daily sample has the same number of variables and they are presented in the same order. Keeping the same data dimension at every sample time is very important as one of the advantages of the SOM clustering methods over k-means algorithm as used in LOESS regression is that the SOM is topology preserving (Lobo 2009). As mentioned in section 3, this means that data “patterns” are preserved in the mapping process and data that have similar patterns in the data matrix are close in the SOM neural map space. Therefore, the use of the ITCOMPSOM method is limited to data that have the same data structure. For ocean moorings, that due to mooring motion sample over depth ranges and have instruments that sample at different frequencies, the data must be gridded in depth and time.

The training phase of the SOM neural classes is done in two steps. The first step uses large neighborhood radii and learning rates to initially classify the data to a general area in the neural map. The second step then fine-tunes this initial map using smaller neighborhood radii and learning rates. In the SOM toolbox documentation these are called rough- and

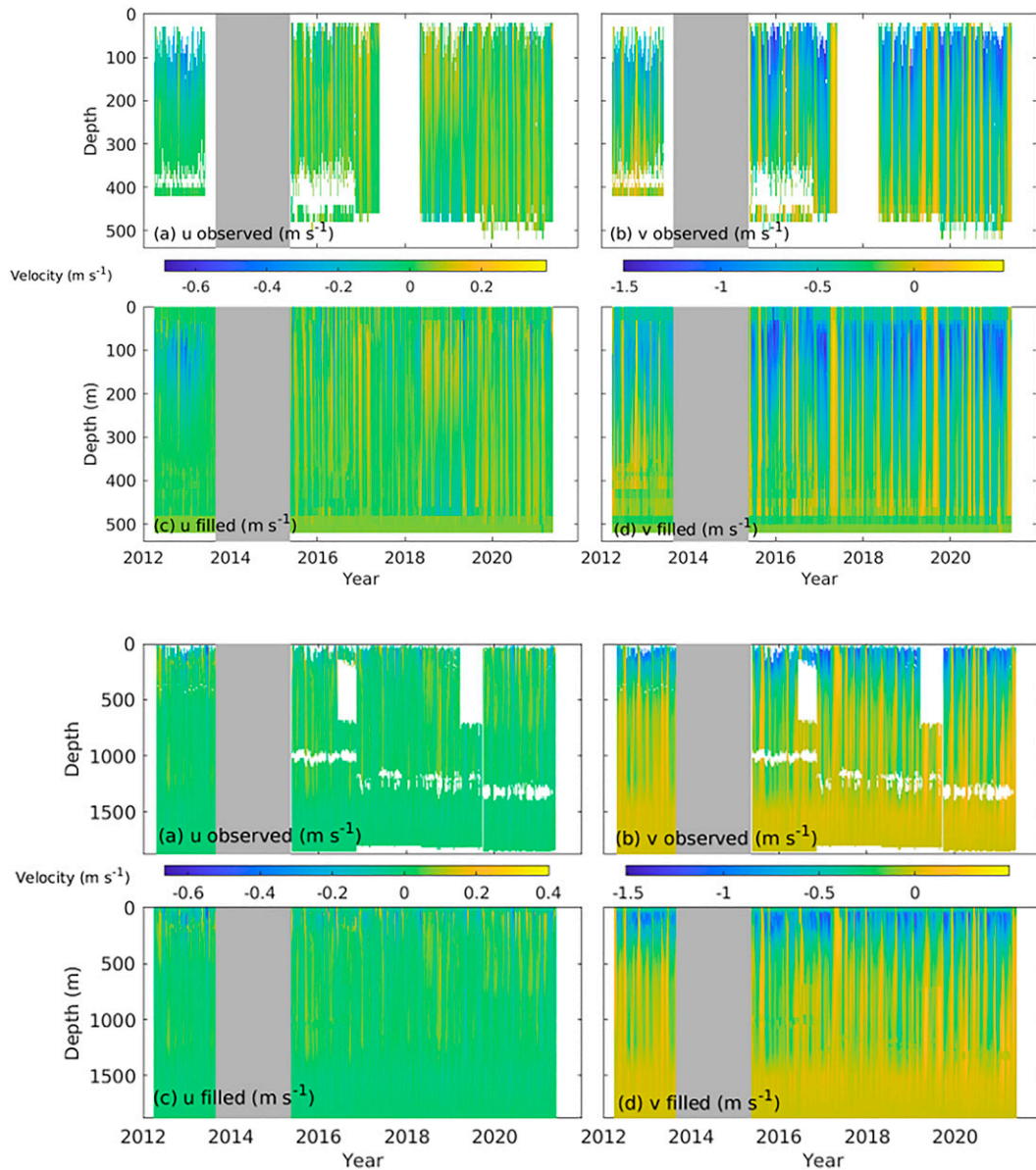


FIG. 9. Comparison of observed and ITCOMP SOM filled u and v velocity (m s^{-1}) for the (top four panels) EAC0500 and (bottom four panels) EAC2000 moorings for observed (a) zonal (u) and (b) meridional (v) velocity and filled missing data (c) zonal (u) and (d) meridional (v) velocity. The 2013–15 period when the mooring array was not deployed is identified by gray.

fine-tuning, respectively. The choice of the neighborhood radii and learning rates can be automatically determined by the `som_make.m` function. Here we use `som_train_struct.m` to obtain the initial training parameters from which we modified to achieve the applied SOM classes. To determine the most appropriate choice of neighborhood radius and learning rates we ran numerous SOM perturbing the initial choices.

It is important to choose an appropriate number of initial neural classes, i_{nc} , as we do not want to generate a SOM that classifies too roughly in the early stages of the iterative process. The best Davies–Bouldin index (Davies and Bouldin

1979), which sets i_{nc} , is determined by iteratively training classes of different sizes on the complete data. The choice of number of iteration n_{it} and final number of classes are considered together as we want to maintain a similar average number of elements in each class as we add new vectors at each iteration.

The best choice of neighborhood radius and learning rates, and initial and final number of SOM classes is determined by careful examination of the \mathbf{U} matrices and number of classifications in each SOM class and their distribution in the two-dimensional space (Lobo 2009; Vesanto et al. 2000). We used

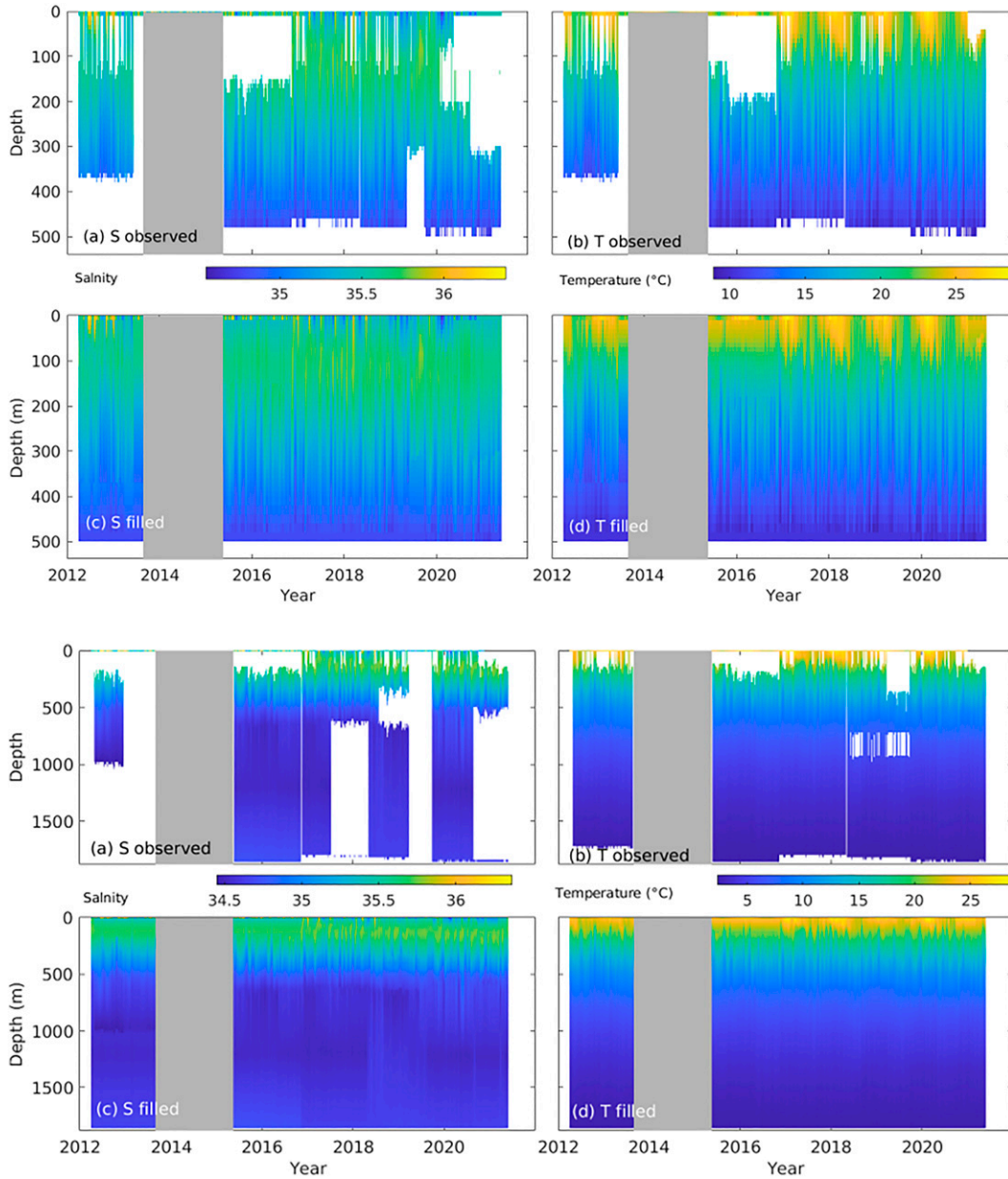


FIG. 10. Comparison of observed and ITCOMPSOM filled salinity (S) and temperature (T , $^{\circ}\text{C}$) for (top four panels) EAC0500 and (bottom four panels) EAC2000 moorings for observed (a) salinity and (b) temperature and filled missing data (c) salinity and (d) temperature. The 2013–15 period when the mooring array was not deployed is identified by gray.

the SOM toolbox diagnostic tools `som_show` and `som_cplane` to visualize these features of the SOM. Determination of the SOM parameters requires the users to run the SOM a number of times to determine the best parameter choices for their data. This can be a timely exercise depending on the size of the training data and may require moderate compute resources to reduce the time taken to adequately explore the parameter space appropriately.

Within the EAC, and many WBC systems, large vertical property gradients are found in the upper ocean (0–1000 m)

(Figs. 2–4). To observe this vertical structure, we instrumented the mooring array to obtain velocity profile data at 4, 8, and 16 m resolution between 30 and 1000 m and temperature and salinity data at vertical resolution of 20 m between 20 and 200 m, at 50 m between 200 and 500 m, and at 100–200 m vertical resolution between 500 and 1000 m. Note that salinity observations were successively improved over time to obtain this resolution. Vertical interpolation to fill missing instrumental observations, particularly at key depths such as near strong vertical velocity, temperature, and salinity gradients

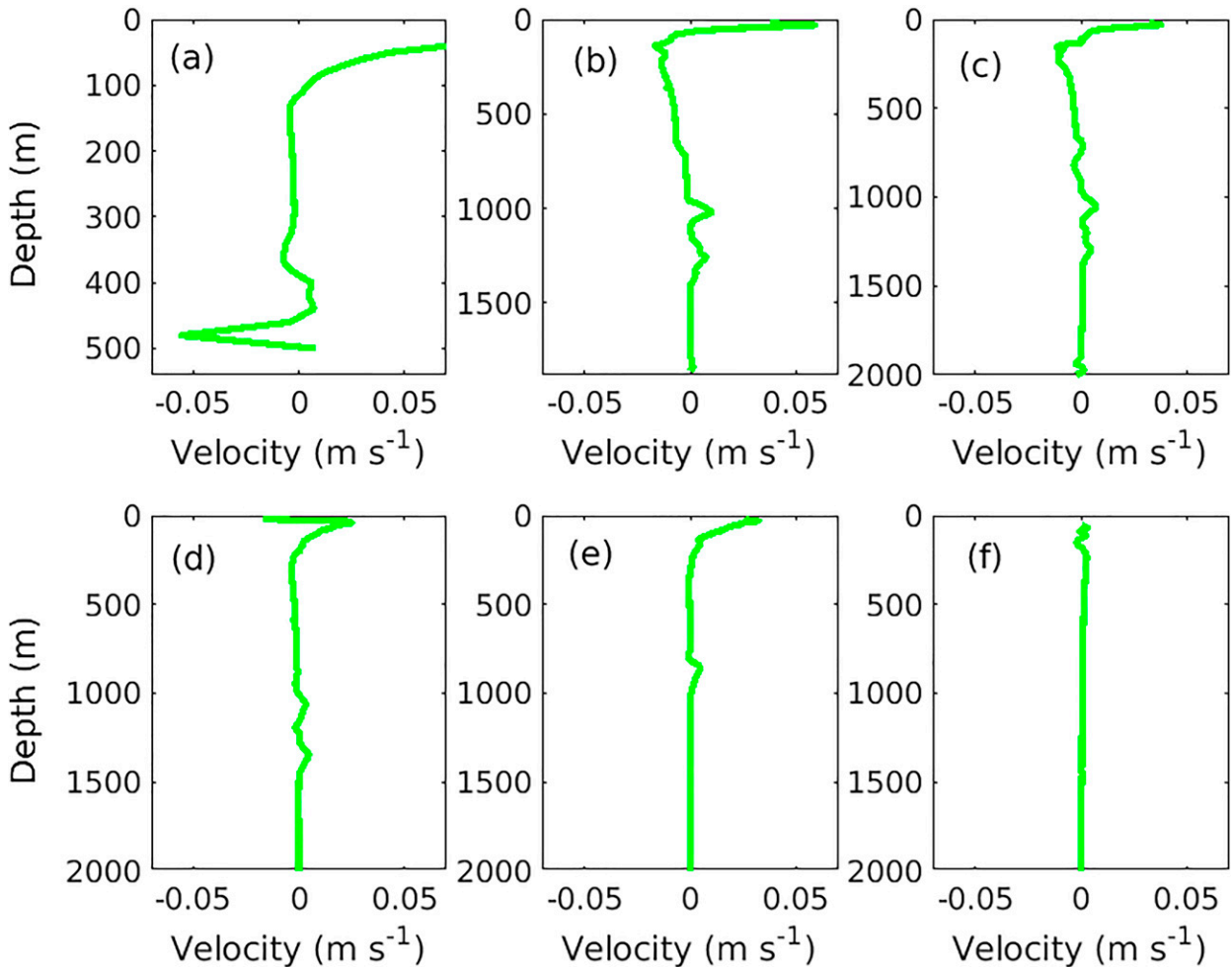


FIG. 11. Residual of time-mean observed minus filled v -velocity profile for (a) EAC0500, (b) EAC2000, (c) EAC3200, (d) EAC4200, (e) EAC4700, and (f) EAC4800 moorings. The profiles are shown for the full depth of the EAC0500 and EAC2000 mooring and upper 2000 m for the remaining moorings.

surrounding the subsurface jet, or fronts, the seasonal and permanent thermocline and halocline, respectively, may produce erroneous and unrealistic vertical property profiles. Thus, while vertical interpolation may be appropriate in some instances to fill missing data gaps, careful examination of the data must be undertaken at each time step prior to its application. We suggest the ITCOMPSOM can be a more effective and efficient method to fill missing data. The SOM assigns each daily vertical data profile into a specific neural class within the two-dimensional neural map. The similarity function that considers both Euclidean (data) distance and correlation between missing and available data is used to choose the best class to fill missing profile data. The mooring vertical profile of data assigned to the class is used to fill missing data in the daily vertical profile. We find that the class mean profile preserves the vertical structure and temporal variability of the time series (Figs. 6c,d, 7, and 8c,d).

With respect to temporal data gaps, Sloyan et al. (2016) found that the velocity integral time scale of the mooring

array, or decorrelation time scale, varies from 4 to 20 days. With the decorrelation time scale being shortest over the continental slope and increasing with distance from the coast and with depth. Thus, it may be appropriate to fill data gaps of less than 20 days using linear interpolation methods. However, larger temporal data gaps should be filled by other methods. Here we find that the ITCOMPSOM filled data preserve the temporal variability of data (Fig. 7), and it is an appropriate method to fill large temporal data gaps.

Given how the SOM assigns each vertical profile to a neural class during the training phase and then uses the mean of profiles assigned to a class to fill data gaps a significant caveat when implementing the method is that the training data should sample a large representative data space and multiple times. Therefore, a long time series is required to best ensure that adequate sampling of the data space is achieved. Temporally long and consistent in situ observations of the ocean have only become available within the last decade due to the

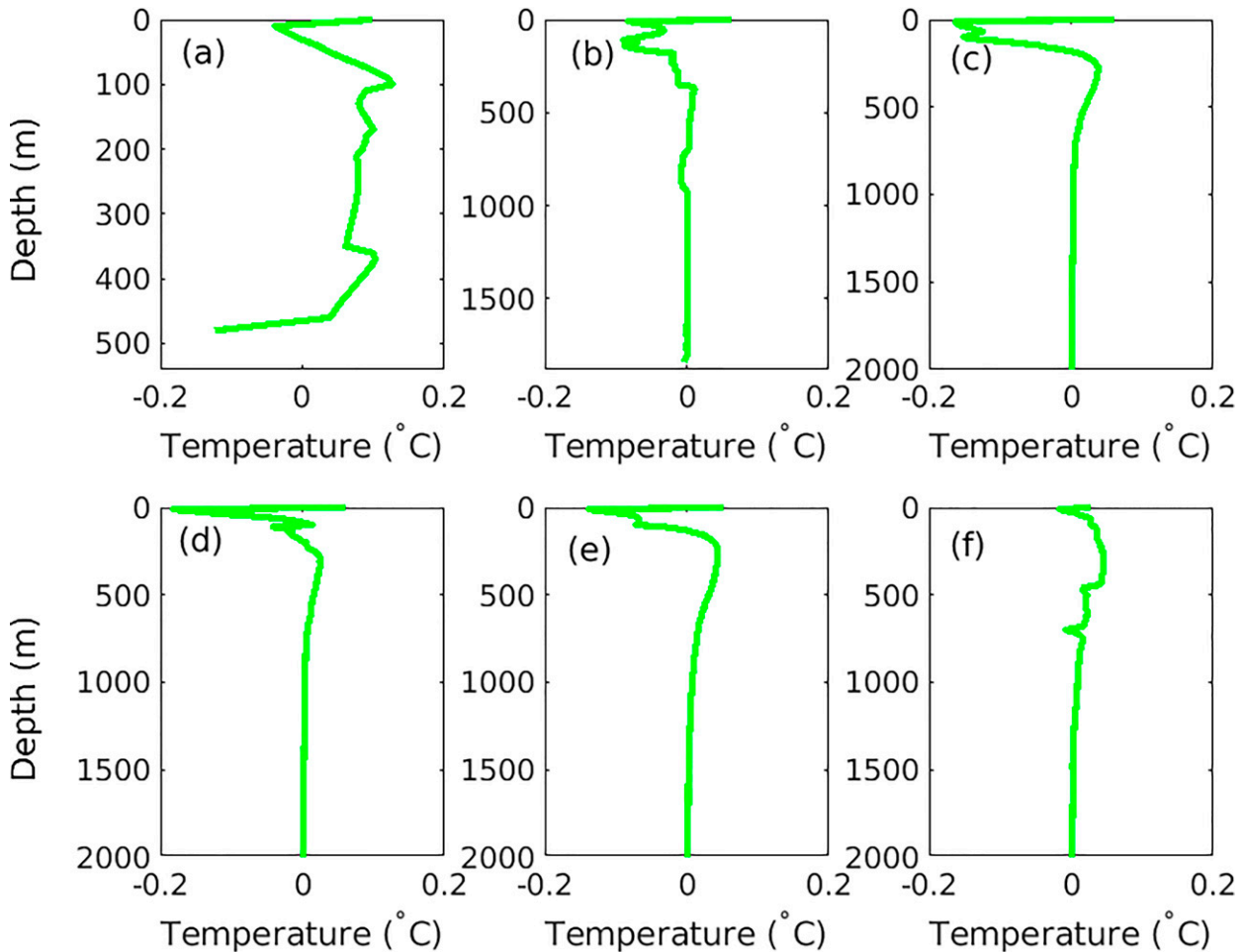


FIG. 12. As in Fig. 11, but for the residual of the time-mean temperature profile.

communities efforts to build a sustained ocean observing system. Within the sustained ocean observing system there are only a few ocean boundary current monitoring systems that have record lengths that are appropriate for machine learning applications. Here we assume that the 7.5 yr time series is of appropriate length.

6. Conclusions

Here we have shown the utility of machine learning, namely, ITCOMPSON, to fill large data gaps in ocean mooring instrumental records. SOM and other machine learning techniques rely on having an appropriate training dataset. The 7.5 years of daily data from the EAC mooring array combined with satellite altimetry, and SSS and SST data are used to train the SOM.

The East Australian Current affects the climate and marine environment from Brisbane to Hobart, an area where more than half of Australia's population reside, which is the site of major coastal infrastructure and is where large agriculture and marine industries operate and significant coastal and

marine biodiversity regions are found (Sloyan et al. 2020). The east Australian shelf and Tasman Sea is a climate change hotspot, warming 4 times faster than the global average (Bindoff et al. 2019). There is increasing evidence that this extreme climate trend may be associated with changes in the magnitude and behavior of the EAC at 26°–28°S (Li et al. 2021; Malan et al. 2021; Kerry et al. 2018; Kerry and Roughan 2020; Sloyan and O'Kane 2015). This ocean warming has strong feedbacks to the frequency and intensity of severe east coast weather. Also, the rapid environmental change is disrupting the marine ecosystems: species ranges are extending or contracting (tropical fish and hard corals are now found near Sydney at latitude 33°S; Booth and Sear 2018), and habitats are changing (kelp forests are disappearing) (Bindoff et al. 2019). These environmental changes are associated with a changing EAC, and carry important risks (and potentially opportunities) for society, agriculture, aquaculture, fisheries, and biodiversity. Similar pressures are found in other WBC regions (Todd et al. 2019).

Well-reasoned climate adaption and mitigation policies, and marine, agriculture, and infrastructure management

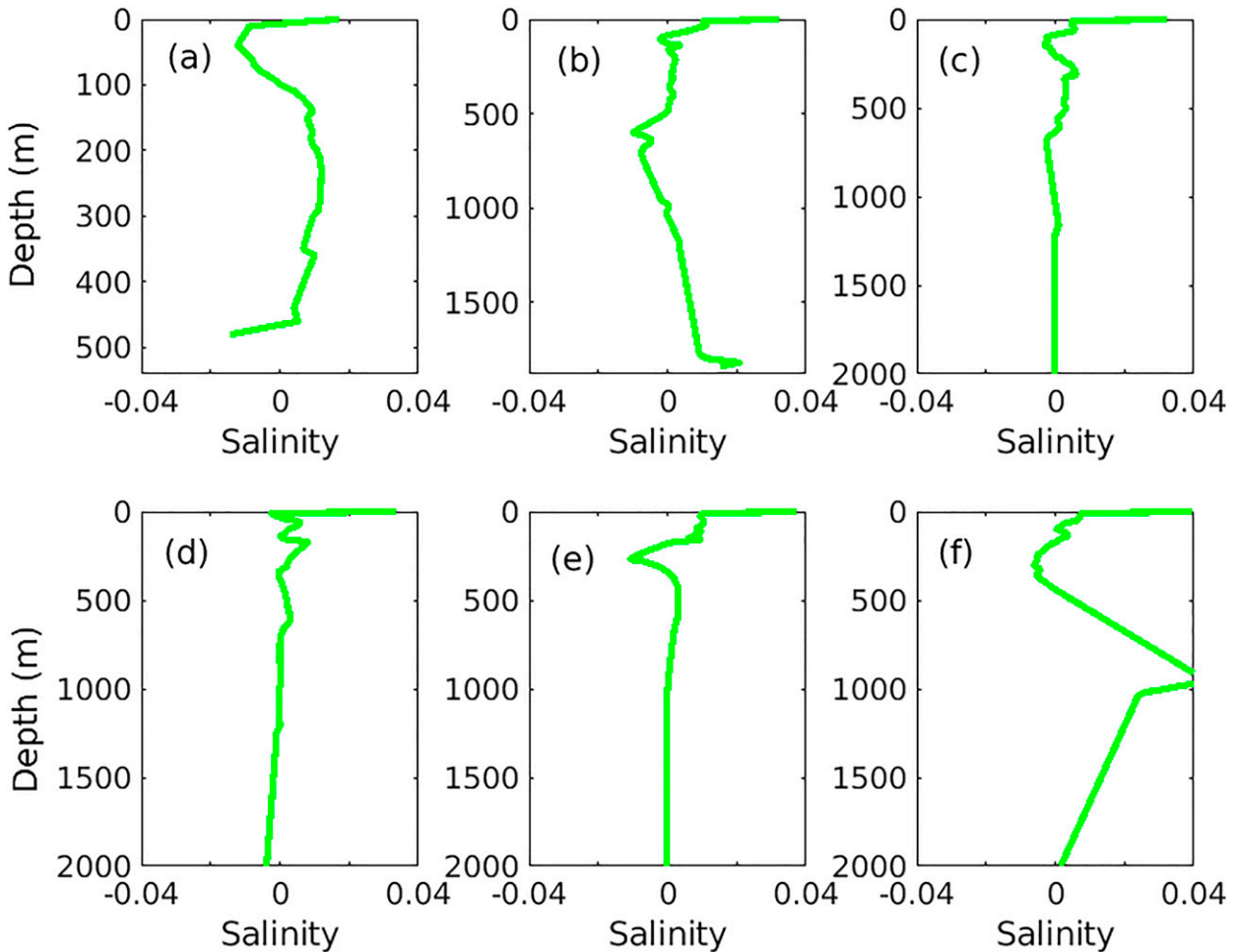


FIG. 13. As in Fig. 11, but for the residual of the time-mean salinity profile.

decisions for the east coast of Australia require high-quality information, including baseline environmental data. The EAC mooring array is one component of the ocean observation system providing some of the temperature, salinity, and current data. From the mooring array data, many potential data users will require consistent (i.e., control volume = 0–1500 m) data products that are derived from the mooring array, such as time series of cross- and along-shelf mass, heat, and salt transport estimates or time series analysis (such as principal component analysis and Fourier analysis) of EAC velocity and transport variability. Prior to production of these time series products gaps in the observational record need to be filled.

The filled EAC mooring data and derived products provide an observational-based time series dataset that can be used to investigate the temporal and spatial variability of the EAC, processes that drive or inhibit shelf–open ocean exchange, assessment of EAC variability, and ocean environmental influences on the marine ecosystem and marine industry productivity. These data products may be added to the suite of ocean data used by a variety of climate and Earth system models to provide more reliable climate predictions for Australia.

Acknowledgments. We thank the CSIRO mooring and instrument team, the MNF for ship time, the R/V *Southern Surveyor* and R/V *Investigator* captains and crew, and the various voyage science parties for their support in the successful maintenance of the EAC mooring array. The EAC mooring array was funded by IMOS and CSIRO Oceans and Atmosphere. BS, CC, and RC were funded by the Center for Southern Hemisphere Oceans Research and CSIRO Oceans and Atmosphere Decadal Climate Forecasting Project (DCFP). BS thanks James Anderson for technical support that enabled the revision of the manuscript while she was working remotely. Comments from Falk Huettmann and one anonymous reviewer improved the manuscript.

Data availability statement. Data were sourced from Australia's Integrated Marine Observing System (IMOS)—IMOS is enabled by the National Collaborative Research Infrastructure Strategy (NCRIS). The IMOS mooring data used are available from the AODN data portal. The compiled EAC hourly mooring data (<https://doi.org/10.25919/bqrf-s872>) and the daily unfilled and SOM filled data product (<https://doi.org/10.25919/fs7p-j741>) are available from the CSIRO data portal.

REFERENCES

- Bindoff, N. L., and Coauthors, 2019: Changing ocean, marine ecosystems, and dependent communities. *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*, H.-O. Pörtner et al., Eds., Cambridge University Press, 447–587, <https://doi.org/10.1017/9781009157964.007>.
- Booth, D. J., and J. Sear, 2018: Coral expansion in Sydney and associated coral-reef fishes. *Coral Reefs*, **37**, 995, <https://doi.org/10.1007/s00338-018-1727-5>.
- Boutin, J., and Coauthors, 2018: New SMOS sea surface salinity with reduced systematic errors and improved variability. *Remote Sens. Environ.*, **214**, 115–134, <https://doi.org/10.1016/j.rse.2018.05.022>.
- , J.-J. Vergely, and D. Khvorostyanov, 2020: SMOS SSS L3 maps generated by CATDS CEC LOCEAN, debias version 7.0. SEANOE, accessed 8 November 2022, <https://doi.org/10.17882/52804>.
- Chapman, C., and A. A. Charantonis, 2017: Reconstruction of subsurface velocities from satellite observations using iterative self-organizing maps. *IEEE Geosci. Remote Sens. Lett.*, **14**, 617–620, <https://doi.org/10.1109/LGRS.2017.2665603>.
- Charantonis, A. A., P. Testor, L. Mortier, F. D'Ortenzio, and S. Thiria, 2015: Completion of a sparse GLIDER database using multi-iterative self-organizing maps (ITCOMP SOM). *Procedia Comput. Sci.*, **51**, 2198–2206, <https://doi.org/10.1016/j.procs.2015.05.496>.
- Cowley, R., 2021: Report on the quality control of the IMOS East Australian Current (EAC) deep water moorings array, version 3.0. CSIRO Oceans and Atmosphere Tech. Rep., 134 pp., <https://doi.org/10.26198/N3XJ-SY16>.
- , 2022a: Report on the quality control of the IMOS East Australian Current (EAC) deep water moorings array, version 1.3. CSIRO Oceans and Atmosphere Tech. Rep., 56 pp., <https://doi.org/10.26198/5r16-xf23>.
- , 2022b: Report on the quality control of the IMOS East Australian Current (EAC) deep water moorings array, version 1.0. CSIRO Oceans and Atmosphere Tech. Rep., 62 pp., <https://doi.org/10.26198/5pzy-cm87>.
- Davies, D. L., and D. W. Bouldin, 1979: A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-1**, 224–227, <https://doi.org/10.1109/TPAMI.1979.4766909>.
- Davis, R. E., and Coauthors, 2019: 100 years of progress in ocean observing systems. *A Century of Progress in Atmospheric and Related Sciences: Celebrating the American Meteorological Society Centennial*, *Meteor. Monogr.*, No 59, Amer. Meteor. Soc., <https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0014.1>.
- Deng, X., D. A. Griffin, K. Ridgway, J. A. Church, W. E. Featherstone, N. J. White, and M. Cahill, 2011: Satellite altimetry for geodetic, oceanographic, and climate studies in the Australian region. *Coastal Altimetry*, S. Vignudelli et al., Eds., Springer, 473–508, https://doi.org/10.1007/978-3-642-12796-0_18.
- Fradkov, A. L., 2020: Early history of machine learning. *IFAC-PapersOnLine*, **53**, 1385–1390, <https://doi.org/10.1016/j.ifacol.2020.12.1888>.
- Frajka-Williams, E., B. I. Moat, D. Smeed, D. Rayner, W. E. Johns, M. O. Baringer, D. L. Volkov, and J. Collins, 2021: Atlantic meridional overturning circulation observed by the RAPID-MOCHA-WBTS (RAPID-Meridional Overturning Circulation and Heat flux Array-Western Boundary Time Series) array at 26°N from 2004 to 2020, version 2020.1. British Oceanographic Data Centre, accessed 1 March 2022, <https://doi.org/10.5285/cc1e34b3-3385-662b-e053-6c86abc03444>.
- Gould, J., B. Sloyan, and M. Visbeck, 2013: In situ ocean observations: A brief history, present status, and future directions. *Ocean Circulation and Climate*, G. Siedler et al., Eds., International Geophysics Series, Vol. 103, Academic Press, 59–81, <https://doi.org/10.1016/B978-0-12-391851-2.00003-9>.
- Gwyther, D. E., C. Kerry, M. Roughan, and S. R. Keating, 2022: Observing system simulation experiments reveal that subsurface temperature observations improve estimates of circulation and heat content in a dynamic western boundary current. *Geosci. Model Dev.*, **15**, 6541–6565, <https://doi.org/10.5194/gmd-15-6541-2022>.
- Hsieh, W. W., and B. Tang, 1998: Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bull. Amer. Meteor. Soc.*, **79**, 1855–1870, [https://doi.org/10.1175/1520-0477\(1998\)079<1855:ANNMTP>2.0.CO;2](https://doi.org/10.1175/1520-0477(1998)079<1855:ANNMTP>2.0.CO;2).
- Huang, B., C. Liu, V. Banzon, E. Freeman, G. Graham, B. Hankins, T. Smith, and H.-M. Zhang, 2021: Improvements of the Daily Optimum Interpolation Sea Surface Temperature (DOISST) version 2.1. *J. Climate*, **34**, 2923–2939, <https://doi.org/10.1175/JCLI-D-20-0166.1>.
- Johns, W. E., T. N. Lee, D. Zhang, R. Zantopp, C.-T. Liu, and Y. Yang, 2001: The Kuroshio east of Taiwan: Moored transport observations from the WOCE PCM-1 array. *J. Phys. Oceanogr.*, **31**, 1031–1053, [https://doi.org/10.1175/1520-0485\(2001\)031<1031:TKEOTM>2.0.CO;2](https://doi.org/10.1175/1520-0485(2001)031<1031:TKEOTM>2.0.CO;2).
- Johnson, E. S., and M. J. McPhaden, 1993: Structure of intraseasonal Kelvin waves in the equatorial Pacific Ocean. *J. Phys. Oceanogr.*, **23**, 608–625, [https://doi.org/10.1175/1520-0485\(1993\)023<0608:SOIKWI>2.0.CO;2](https://doi.org/10.1175/1520-0485(1993)023<0608:SOIKWI>2.0.CO;2).
- Joseph, V. R., 2022: Optimal ratio for data splitting. *Stat. Anal. Data Min.*, **15**, 531–538, <https://doi.org/10.1002/sam.11583>.
- Kanzow, T., U. Send, W. Zenk, A. D. Chave, and M. Rhein, 2006: Monitoring the integrated deep meridional flow in the tropical North Atlantic: Long-term performance of a geostrophic array. *Deep-Sea Res. I*, **53**, 528–546, <https://doi.org/10.1016/j.dsr.2005.12.007>.
- Kerry, C., and M. Roughan, 2020: Downstream evolution of the East Australian Current System: Mean flow, seasonal and intra-annual variability. *J. Geophys. Res. Oceans*, **125**, e2019JC015227, <https://doi.org/10.1029/2019JC015227>.
- , —, and B. Powell, 2018: Observation impact in a regional reanalysis of the East Australian Current System. *J. Geophys. Res. Oceans*, **123**, 7511–7528, <https://doi.org/10.1029/2017JC013685>.
- Kohonen, T., 2001: *Self-Organizing Maps*. 3rd ed. Vol. 30, Springer, 502 pp., <https://doi.org/10.1007/978-3-642-56927-2>.
- , 2013: Essentials of the self-organizing map. *Neural Networks*, **37**, 52–65, <https://doi.org/10.1016/j.neunet.2012.09.018>.
- Li, J., M. Roughan, and C. Kerry, 2021: Dynamics of interannual eddy kinetic energy modulations in a western boundary current. *Geophys. Res. Lett.*, **48**, e2021GL094115, <https://doi.org/10.1029/2021GL094115>.
- Li, X., and Coauthors, 2020: Moored observations of transport and variability of Halmahera Sea currents. *J. Phys. Oceanogr.*, **50**, 471–488, <https://doi.org/10.1175/JPO-D-19-0109.1>.
- Lobo, V. J. A. S., 2009: Application of self-organising maps to the maritime environment. *Information Fusion and Geographic Information Systems*, V. V. Popovich et al., Eds., Springer-Verlag, 19–36, https://doi.org/10.1007/978-3-642-00304-2_2.
- Lovell, J., and R. Cowley, 2022a: Report on the quality control of the IMOS East Australian Current (EAC) deep water

- moorings array, version 3.2. CSIRO Oceans and Atmosphere Tech. Rep., 142 pp., <https://doi.org/10.26198/5d3fb95821dda>.
- , and —, 2022b: Report on the quality control of the IMOS East Australian Current (EAC) deep water moorings array, version 1.2. CSIRO Oceans and Atmosphere Tech. Rep., 162 pp., <https://doi.org/10.26198/Sec1df4b25cca>.
- Malan, N., M. Roughan, and C. Kerry, 2021: The rate of coastal temperature rise adjacent to a warming western boundary current is nonuniform with latitude. *Geophys. Res. Lett.*, **48**, e2020GL090751, <https://doi.org/10.1029/2020GL090751>.
- McMonigal, K., L. M. Beal, S. Elipot, K. L. Gunn, J. Hermes, T. Morris, and A. Houk, 2020: The impact of meanders, deepening and broadening, and seasonality on Agulhas Current temperature variability. *J. Phys. Oceanogr.*, **50**, 3529–3544, <https://doi.org/10.1175/JPO-D-20-0018.1>.
- Molnar, C., G. Casalicchio, and B. Bischl, 2020: Interpretable machine learning—A brief history, state-of-the-art and challenges. *20th Joint European Conf. on Machine Learning and Knowledge Discovery in Databases*, Online, ECML PKDD, 417–431.
- Oke, P. R., and Coauthors, 2019: Revisiting the circulation of the East Australian Current: Its path, separation, and eddy field. *Prog. Oceanogr.*, **176**, 102139, <https://doi.org/10.1016/j.pocan.2019.102139>.
- Palacz, A. P., J. Pearlman, S. Simmons, K. Hill, P. Miloslavich, and M. Telszewski, 2017: Report of the Workshop on the Implementation of Multi-Disciplinary Sustained Ocean Observations (IMSOO). GOOS Tech. Rep. 223, 91 pp., <https://www.gooscean.org/imsoo-report>.
- Pedlosky, J., 1996: *Ocean Circulation Theory*. Springer, 456 pp., https://doi.org/10.1007/978-3-662-03204-6_2.
- Pepler, A. S., L. V. Alexander, J. P. Evans, and S. C. Sherwood, 2016: The influence of local sea surface temperatures on Australian east coast cyclones. *J. Geophys. Res. Atmos.*, **121**, 13 352–13 363, <https://doi.org/10.1002/2016JD025495>.
- Puissant, A., R. El Hourany, A. A. Charantonis, C. Bowler, and S. Thiria, 2021: Inversion of phytoplankton pigment vertical profiles from satellite data using machine learning. *Remote Sens.*, **13**, 1445, <https://doi.org/10.3390/rs13081445>.
- Ridgway, K. R., and J. S. Godfrey, 1997: Seasonal cycle of the East Australian Current. *J. Geophys. Res.*, **102**, 22 921–22 936, <https://doi.org/10.1029/97JC00227>.
- Send, U., R. Davis, J. Fischer, S. Imawaki, W. Kessler, and C. Meinen, 2010: A global boundary current circulation observing network. *Proc. OceanObs'09: Sustained Ocean Observations and Information for Society*, Venice, Italy, ESA, <https://doi.org/10.5270/OceanObs09.cwp.78>.
- Sloyan, B. M., and T. J. O’Kane, 2015: Drivers of decadal variability in the Tasman Sea. *J. Geophys. Res. Oceans*, **120**, 3193–3210, <https://doi.org/10.1002/2014JC010550>.
- , and R. Cowley, 2022: East Australian Current individual mooring gridded product-hourly and 10-20m depth gridded, version 4. CSIRO Data Collection, accessed 16 August 2022, <https://doi.org/10.25919/xkxg-zyl4>.
- , K. R. Ridgway, and R. Cowley, 2016: The East Australian Current and property transport at 27°S from 2012 to 2013. *J. Phys. Oceanogr.*, **46**, 993–1008, <https://doi.org/10.1175/JPO-D-15-0052.1>.
- , M. Cahill, M. Roughan, and K. Ridgway, 2020: East Australian Current variability. State and trends of Australia’s ocean, Integrated Marine Observing System Rep., 1.3.1–1.3.5, <https://doi.org/10.26198/5e16a23f49e75>.
- , R. Cowley, and C. Chapman, 2021: East Australian Current individual mooring gridded product-daily and 10 m depth gridded, version 8. CSIRO Data Collection, accessed 16 August 2022, <https://doi.org/10.25919/a8j3-zh92>.
- Sonnenwald, M., R. Lguensat, D. C. Jones, P. D. Dueben, J. Brajard, and V. Balaji, 2021: Bridging observations, theory and numerical simulation of the ocean using machine learning. *Environ. Res. Lett.*, **16**, 073008, <https://doi.org/10.1088/1748-9326/ac0eb0>.
- Sprintall, J., S. E. Wijffels, R. Molcard, and I. Jaya, 2009: Direct estimates of the Indonesian Throughflow entering the Indian Ocean: 2004–2006. *J. Geophys. Res.*, **114**, C07001, <https://doi.org/10.1029/2008JC005257>.
- Suthers, I. M., and Coauthors, 2011: The strengthening East Australian Current, its eddies and biological effects—An introduction and overview. *Deep-Sea Res. II*, **58**, 538–546, <https://doi.org/10.1016/j.dsr2.2010.09.029>.
- Talley, L. D., G. L. Pickard, W. J. Emery, and J. H. Swift, 2011: *Descriptive Physical Oceanography: An Introduction*. 6th ed. Elsevier, 560 pp.
- Todd, R. E., and Coauthors, 2019: Global perspectives on observing ocean boundary current systems. *Front. Mar. Sci.*, **6**, 423, <https://doi.org/10.3389/fmars.2019.00423>.
- Vesanto, J., J. Himberg, E. Alhoniemi, and J. Parhankangas, 2000: Self-organizing map in MATLAB: The SOM toolbox. Helsinki University of Technology Tech. Rep., 59 pp.
- Wang, Y., M. J. McPhaden, P. Freitag, and C. Fey, 2015: Moored acoustic Doppler current profiler time series in the central equatorial Indian Ocean. NOAA Tech. Memo. OAR PMEL-146, 23 pp, <https://doi.org/10.7289/V5HX19NP>.
- Wood, J. E., A. Schaeffer, M. Roughan, and P. M. Tate, 2016: Seasonal variability in the continental shelf waters off southeastern Australia: Fact or fiction? *Cont. Shelf Res.*, **112**, 92–103, <https://doi.org/10.1016/j.csr.2015.11.006>.