Application of variational autoencoder-based clustering for geophysical fluid circulations with a small sample size

Kunihiro Aoki, Hideyuki Nakano, Nariaki Hirose, Norihisa Usui,
Kei Sakamoto, Takahiro Toyoda, Shogo Urakawa, Yuma Kawakami
(*Meteorological Research Institute, Japan Meteorological Agency*)

AI-TT Workshop, Montreal, Canada
13-14 Apr 2026

HABITABLE JAPAN
HOTSPOT III

JGR | Machine Learning and Computation

Research Article | Open Access

Application of Variational Autoencoder-Based Clustering for Geophysical Fluid Circulations With Small Sample Size

Kunihiro Aoki, Hideyuki Nakano, Nariaki Hirose, Norihisa Usui, Kei Sakamoto, Takahiro Toyoda, Shogo Urakawa, Yuma Kawakami

First published: 17 March 2026 | https://doi.org/10.1029/2025JH001051 |

Introduction

JGR paper

Clustering analysis is a fundamental approach in weather, climate, and oceanic studies.

Why clustering?:

- **Research:**
  We can identify climate regimes to determine the underlying dynamics[e.g. Kimoto & Ghil, 1993b,a; Hannachi & Iqbal, 2019; Aoki et al., 2020; Babanov et al. 2023].
- **Operational forecat:**
  We can summarize huge ensemble forecasts for considering possible scenario[e.g. Ferranti & Corti, 2011; Ferranti et al., 2014].

Introduction

JGR paper

The Challenge:

- Ocean data is too complex for raw analysis, and thus, we need a reasonably compressed feature space to find patterns.

The Gap:

- Classical methods like PCA are linear and simplifies the features too much [Stratus et al 2007; Dawson et al. 2012; Aoki et al. 2020].
- While Neural Networks give us better feature representations, most methods stick a classifier (like, K-means) on the end[e.g., Song et al. 2013; Kurihana et al. 2014; Jiang et al. 2017].

  → *This allows breaking the internal consistency of the model.*

The Solution:

- We may seek a unified framework, requiring no external classifiers.
- The Prasad's approach based on the variational autoencoder delivers this "End-to-end" unity[Prasad et al 2020].

Introduction:

JGR paper

The Problems:

- Any clustering algorithms lack an intrinsic metric to determine the number of clusters.
- Deep learning struggles with small sample sizes.

  — Small-sample-size problem is a common issue for long-time scale or strict conditional constraints in the climate data.
- Data augmentation may overcome this problem, but, the standard methods ignore the geometrical constrains of the geophysical fluid circulations

Pupose of this study:

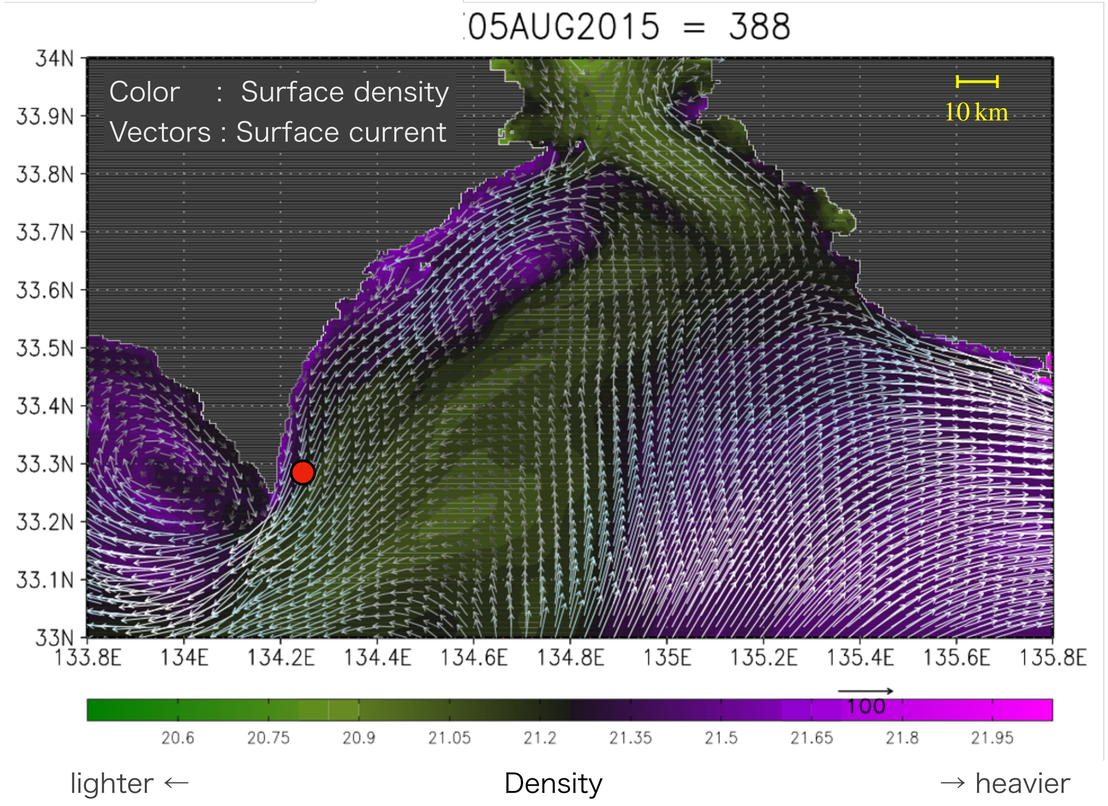
- To propose a physics-aware data augmentation technique.
- To provide a total method including this augmentation to function the Prasad's VAE clustering algorithm for the geophygical fluid circulations

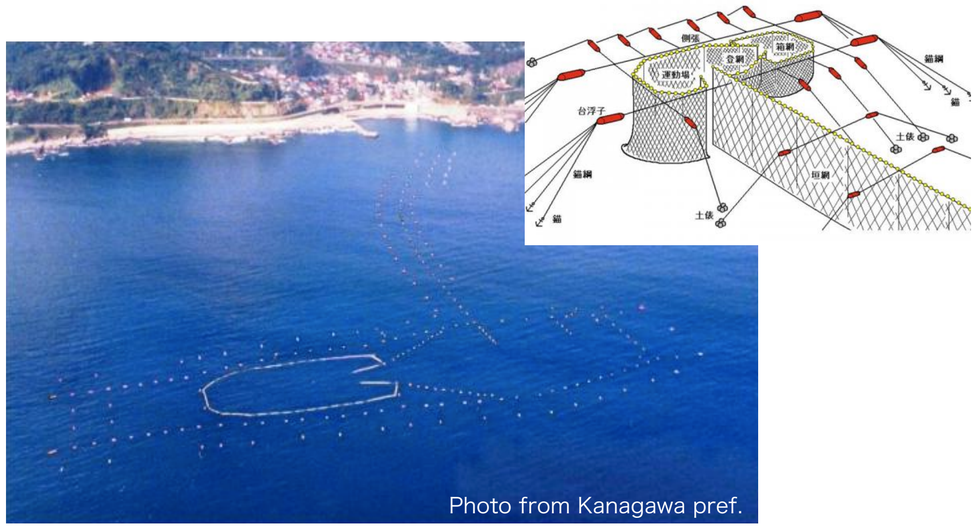
The target phenomenon:

Kyucho: Sporadic strong current event along a coast.

Introduction: What's Kyucho?

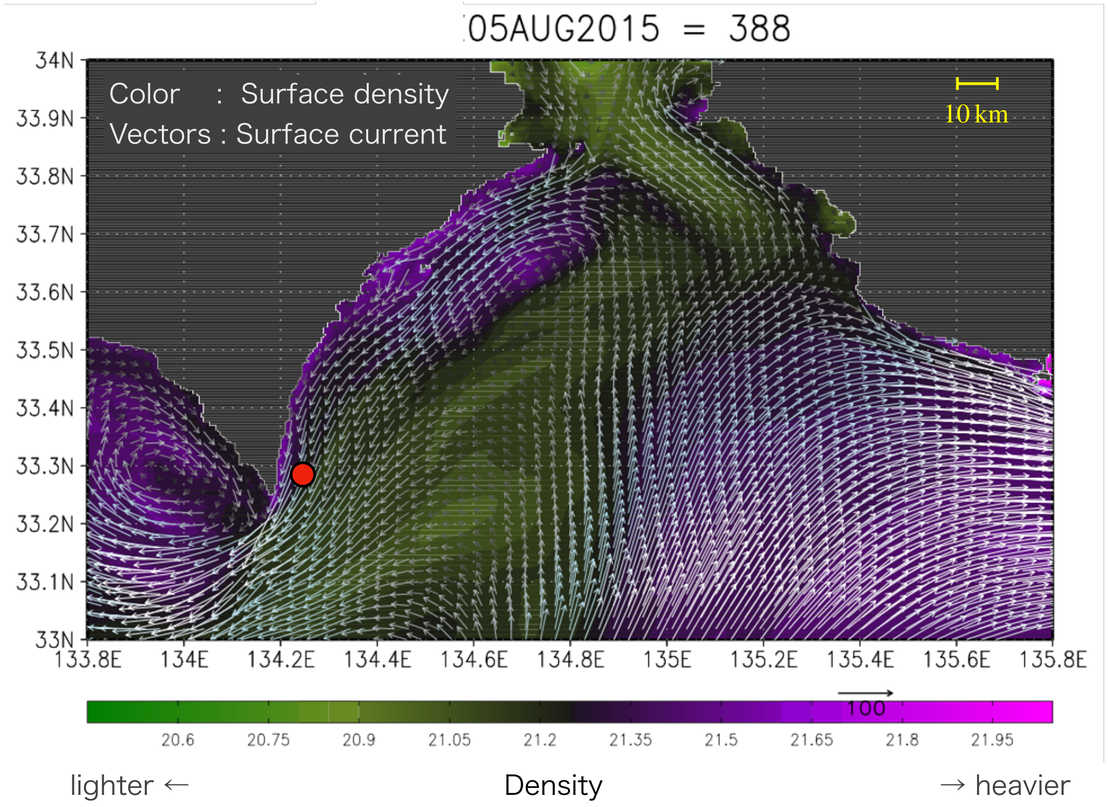
JGR paper

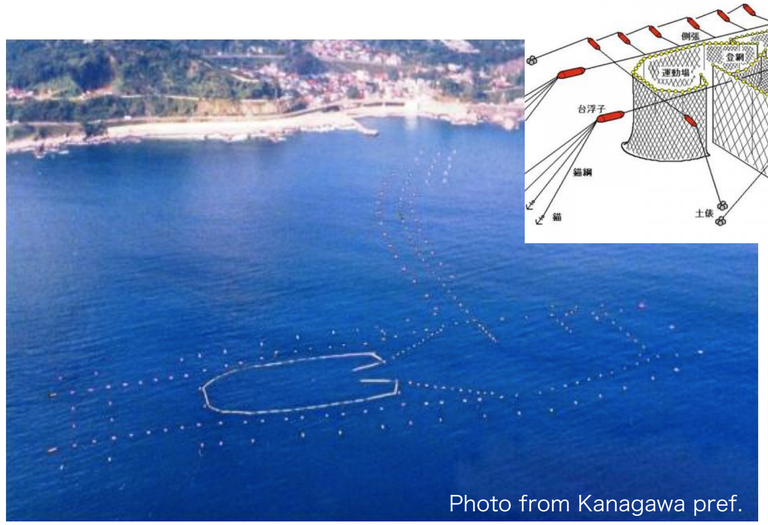




Introduction: What's Kyucho?

JGR paper





surface gravity current

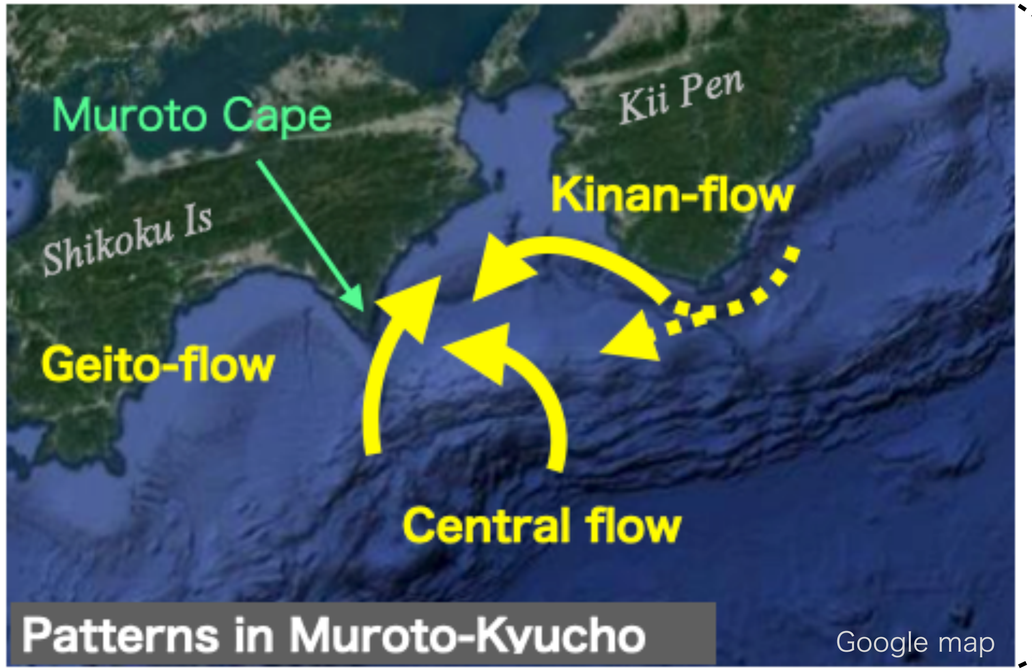
coastal-trapped waves

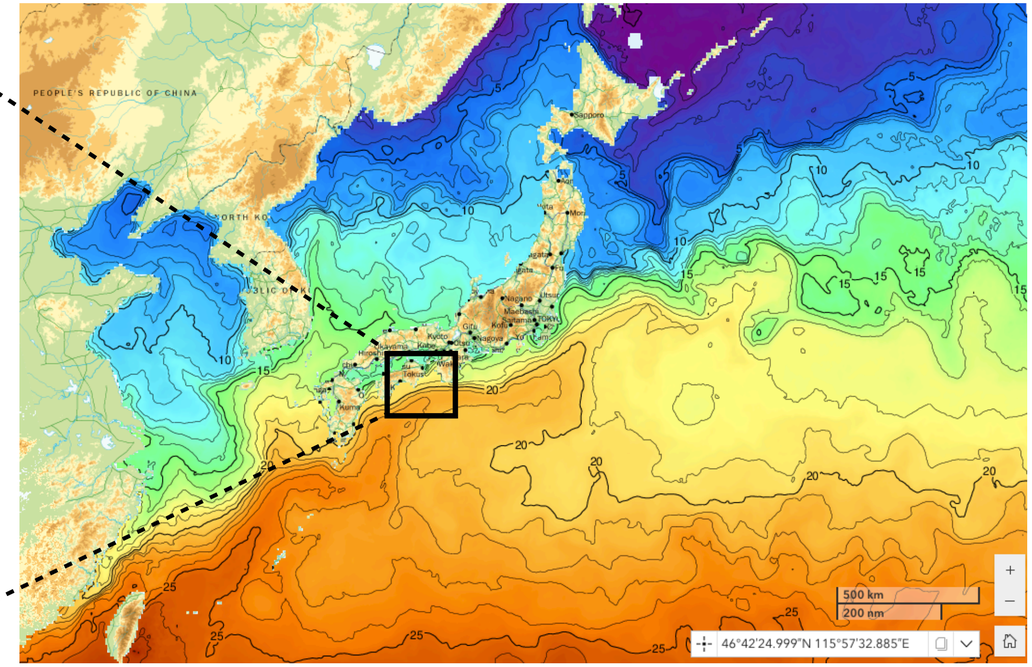
frontal waves *(from Western Boundary Current)*

Introduction: What's Kyucho?

JGR paper

Human-identified flow patterns



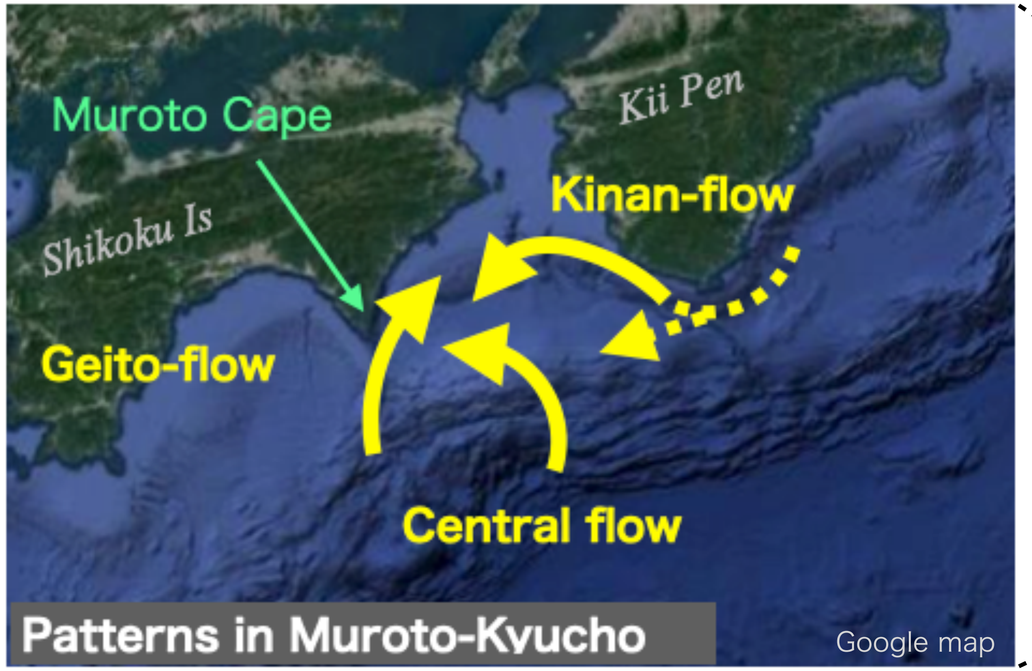


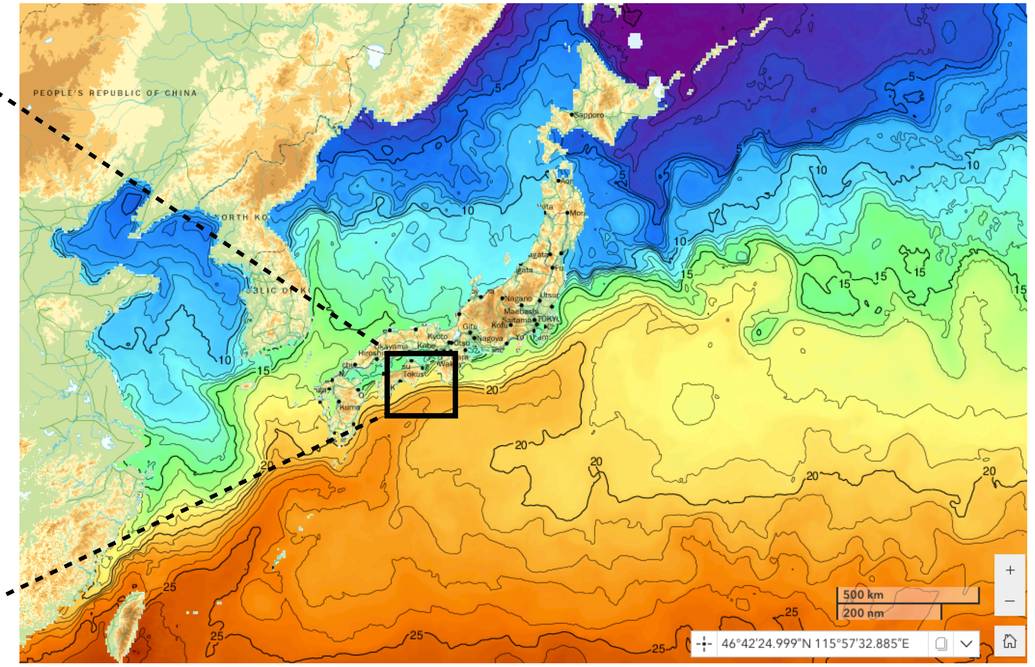
Kochi Prefectural Fisheries Experimental Station

Introduction: What's Kyucho?

JGR paper

Human-identified flow patterns





Kochi Prefectural Fisheries Experimental Station

This study aims to

extract the flow patterns in Kyucho from a long-term reanalysis ocean data

Data: Source

JGR paper

FORA-JPN60

Ocean model	MRI.COM v5.0 ‣ Horizontal resolutions - JPN:2km, NP:10km, GLB:100km ‣ Online two-way nesting ‣ Including tide and pressure adjustment
Forcings	Atmosheric forcings: JRA-3Q (3 hours) River runoff: JRA-55 + JMA-RI
Data assimilation	MOVE_V4 ‣ Ocean: 4D-Var + IAU downscaling ‣ Sea ice: Nudging
Observation data	SST（COBE-SST2，MGDSST） Satelite-based SSH anomaly ‣ Along track sea level anomaly (14 satelites) ‣ Removing non-steric height In-situ temperature and salinity ‣ Basically EN4 but multiple dataset are added Sea ice concentration ‣ SSMI, SSMIS ‣ Okhotsk sea ice analysis
Period	1 Jan 1960 — 31 Dec 2020

(from Usui et al. 2026, J.Oceanogr.)

Data: Source

JGR paper

FORA-JPN60

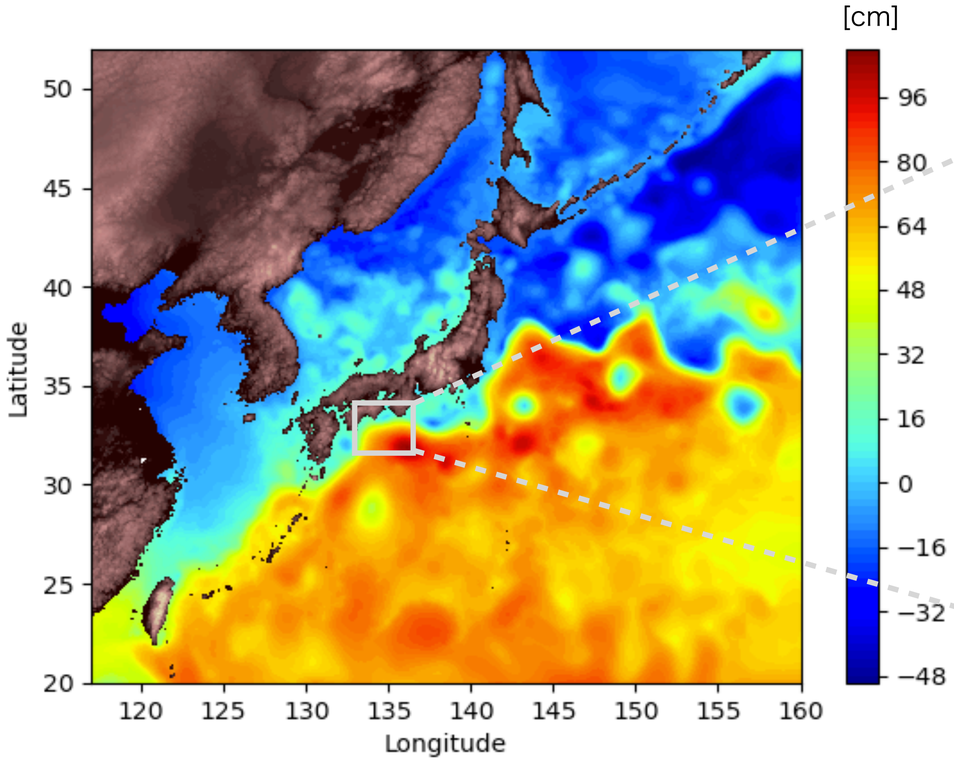
Ocean model	MRI.COM v5.0 ‣ Horizontal resolutions - JPN:2km, NP:10km, GLB:100km ‣ Online two-way nesting ‣ Including tide and pressure adjustment
Forcings	Atmosheric forcings: JRA-3Q (3 hours) River runoff: JRA-55 + JMA-RI
Data assimilation	MOVE_V4 ‣ Ocean: 4D-Var + IAU downscaling ‣ Sea ice: Nudging
Observation data	SST（COBE-SST2，MGDSST） Satelite-based SSH anomaly ‣ Along track sea level anomaly (14 satelites) ‣ Removing non-steric height In-situ temperature and salinity ‣ Basically EN4 but multiple dataset are added Sea ice concentration ‣ SSMI, SSMIS ‣ Okhotsk sea ice analysis
Period	1 Jan 1960 — 31 Dec 2020

(from Usui et al. 2026, J.Oceanogr.)

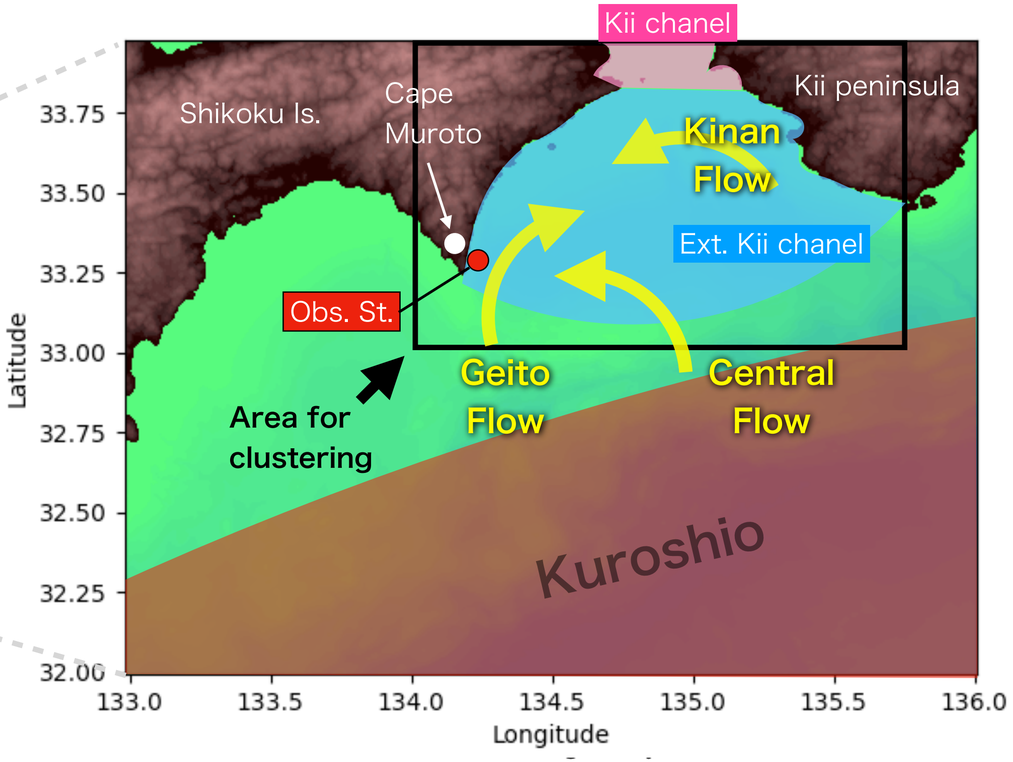
Data: Target region

JGR paper

a) Model domain



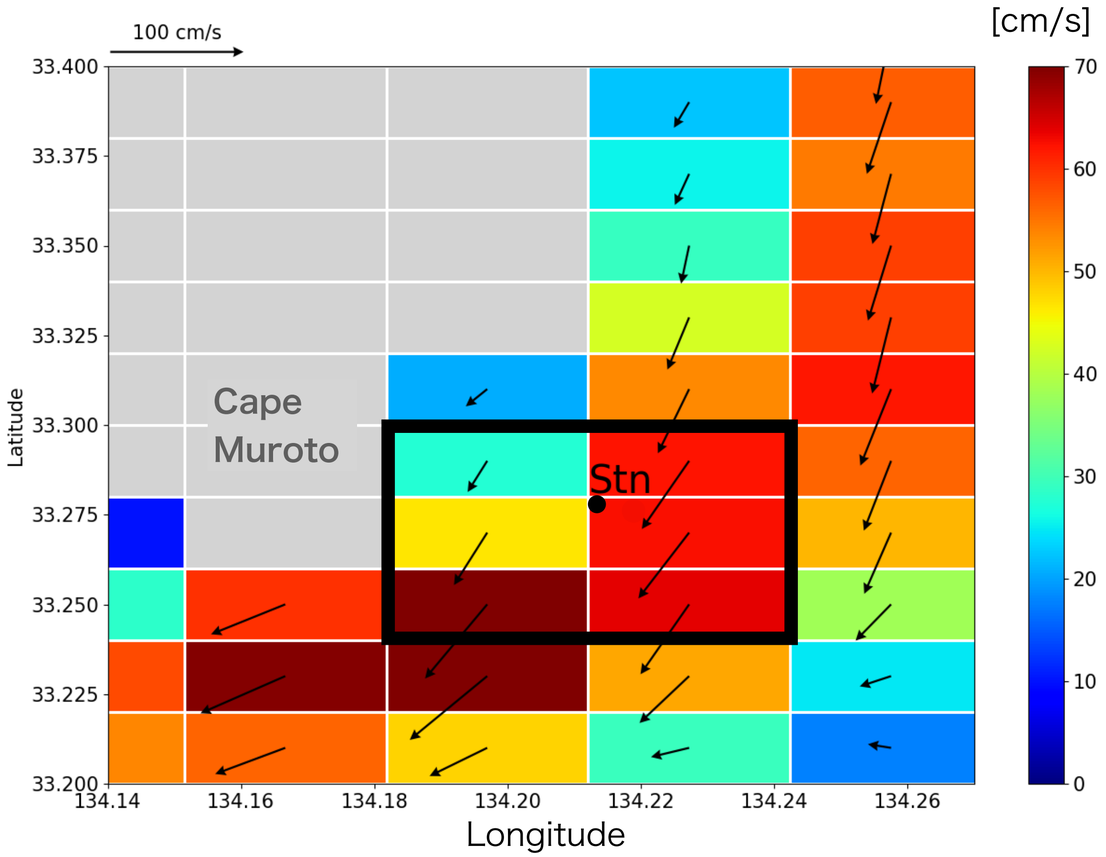
b) Around Cape Muroto

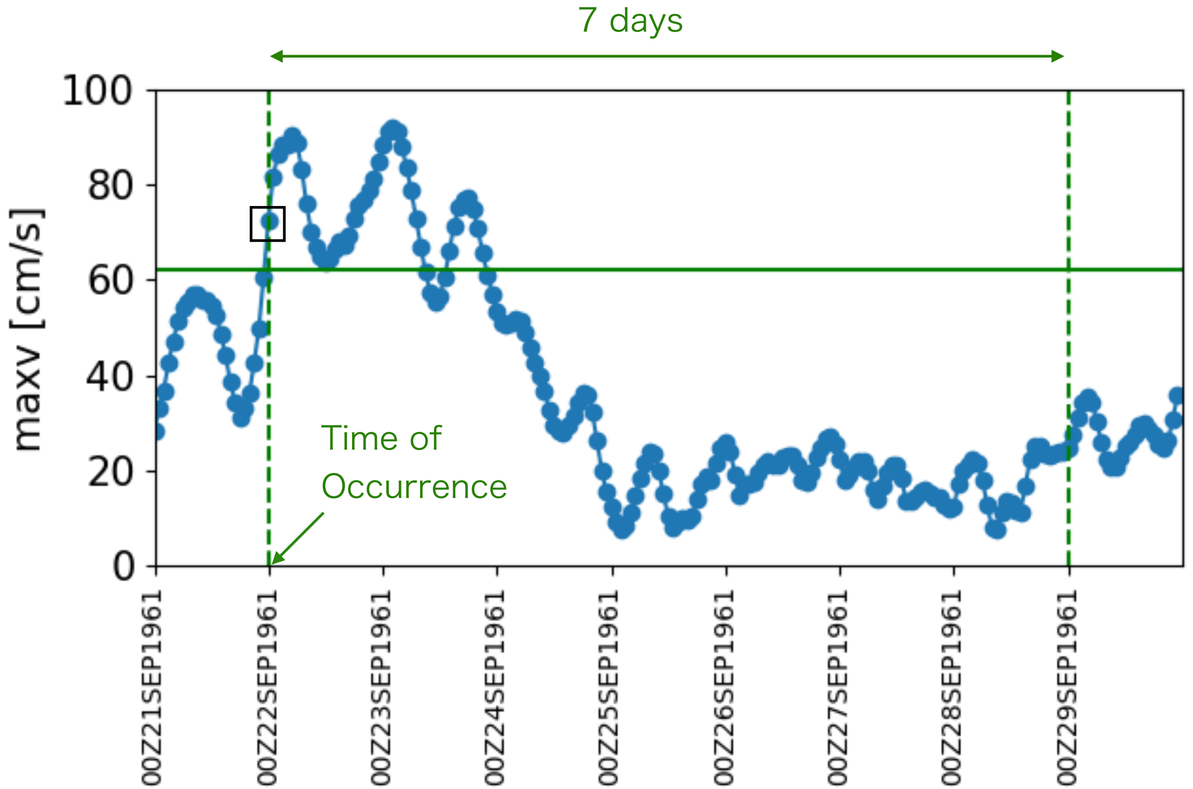


Time extent, resolution	1960–2019, 1hour
Zonal extent, resolution	134.0E–135.7E, ∼2km
Meridional extent, resolution	33N–34.1N, ∼2km
Variables	Surface horizontal velocities

Data: Detection of Kyucho

JGR paper





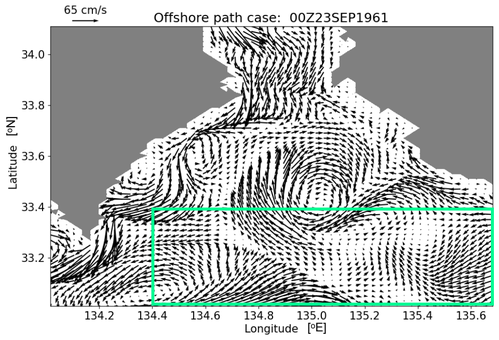
Data: Detection of Kyucho

JGR paper

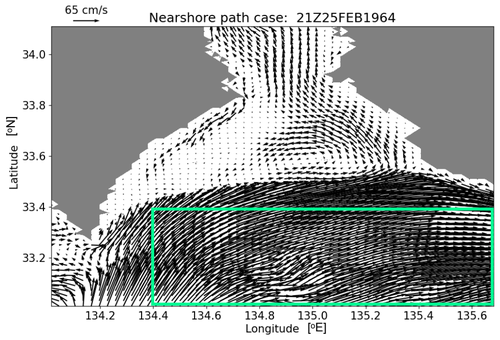
Detected Kyucho:

$N = 721$

Offshore path case: $N = 197$



Nearshore path case: $N = 524$



Training set:
$N =$ **164**

Validation set:
$N = 33$

Very small !!

Training & Validation sets

Method: VAE clustering

JGR paper

Variational inference is performed by maximizing the following ELBO:

$$ELBO = \mathbb{E}_{q,p_D} \log \frac{p_{\beta,\theta}(\mathbf{x}, \mathbf{z}, k)}{q_{\omega,\psi}(\mathbf{z}, k \mid \mathbf{x})}$$

$$\left(\mathbb{E}_{q,p_D} : \text{ Expectation wrt } q_{\omega,\psi}(\mathbf{z}, k \mid \mathbf{x}) p_D(\mathbf{x}) \right)$$

Models

$$p_{\beta,\theta}(\mathbf{x}, \mathbf{z}, k) = p_\theta(\mathbf{x} \mid \mathbf{z}) p_\beta(\mathbf{z} \mid k) p(k)$$

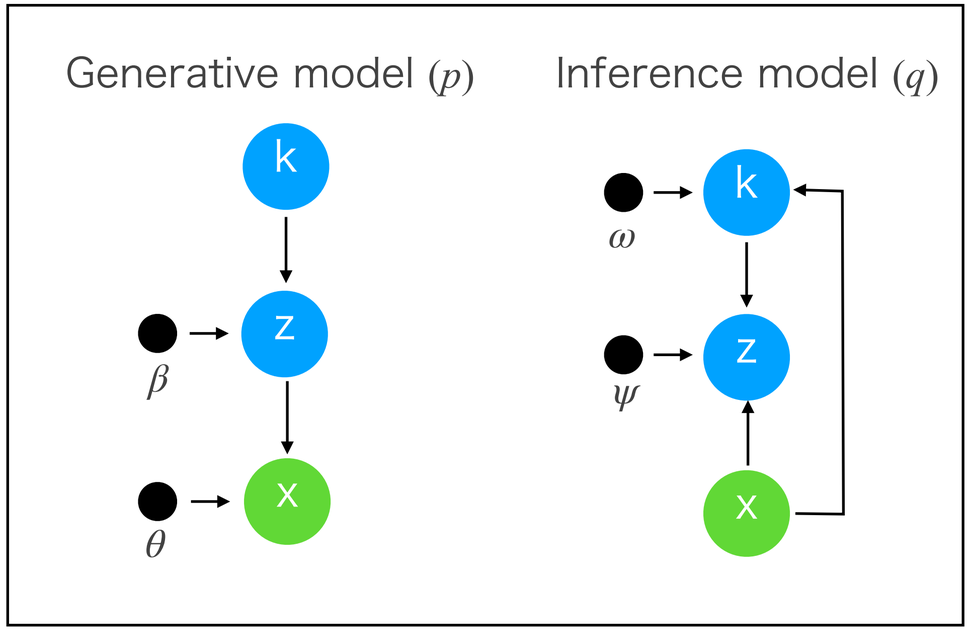
$$q_{\omega,\psi}(\mathbf{z}, k \mid \mathbf{x}) = q_\psi(\mathbf{z} \mid k, \mathbf{x}) q_\omega(k \mid \mathbf{x})$$

← Bayes's decomposition thorem

Assum. Mixture distribution

$$\tilde{p}(x, z) \begin{cases} \equiv \Sigma_k p(x, z, k) \\ = p(x \mid z) p(z) \end{cases}$$

↙ also $p(z) = \Sigma_k p(z \mid k) p(k)$



$$ELBO = \mathbb{E}_{q,p_D}[\ln p_\theta(\mathbf{x} \mid \mathbf{z})] - \underbrace{\mathbb{E}_{p_D} KL(q_\omega(k \mid \mathbf{x}) \| p(k))}_{\gamma_1} - \underbrace{\mathbb{E}_{q_\omega(k|\mathbf{x})p_D}[KL(q_\psi(\mathbf{z} \mid k, \mathbf{x}) \| p_\beta(\mathbf{z} \mid k))]}_{\gamma_2}$$

$$\left(KL(f(x) \| g(x)) := -\sum_x f(x) \ln \frac{g(x)}{f(x)} \right)$$

$$p(k) = Cat(\boldsymbol{\pi})$$ ← Generally, uniform

$$p_\beta(\mathbf{z} \mid k) = \mathcal{N}(\mathbf{z} \mid \mu_\beta(\mathbf{e}_k), diag(\sigma^2_\beta(\mathbf{e}_k)))$$

$$p_\theta(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(\mathbf{x} \mid \mu_\theta(\mathbf{z}), diag(\sigma^2_\theta(\mathbf{z})))$$

Same goes for q, except for the parameters, ω, ψ.

(Prasad et al., 2020)

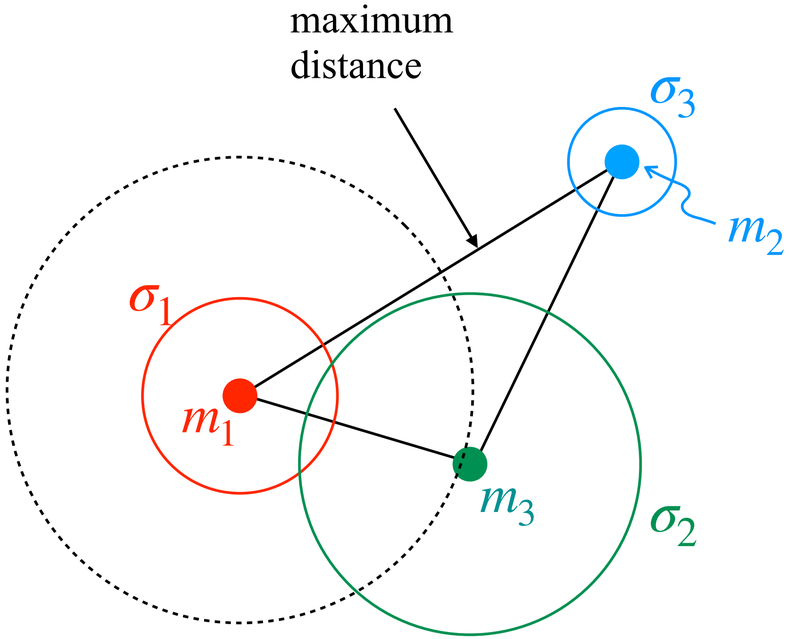
Method application

JGR paper

Number of clusters:

Determining the optimal number of clusters is one of a fundamental problem in any cluster analysis.

- The **Silhouette Score** seems to be the best method [Arbelaitz et al. 2013], but has some limitation under a specific data structure [Liu et al. 2010], and requires the original samples to its calculation.
- So, we provide a “short-cut” approach using **classifiability conditions** based on cluster centroids and variances in the latent space to evaluate the separation between the clusters against their standard deviations. → See our paper



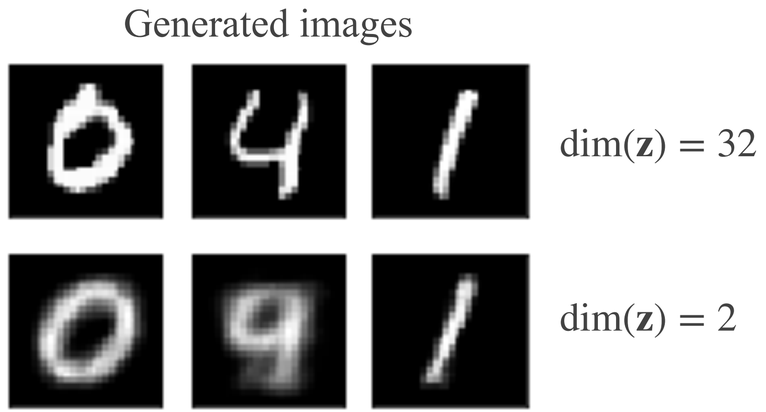
Method application

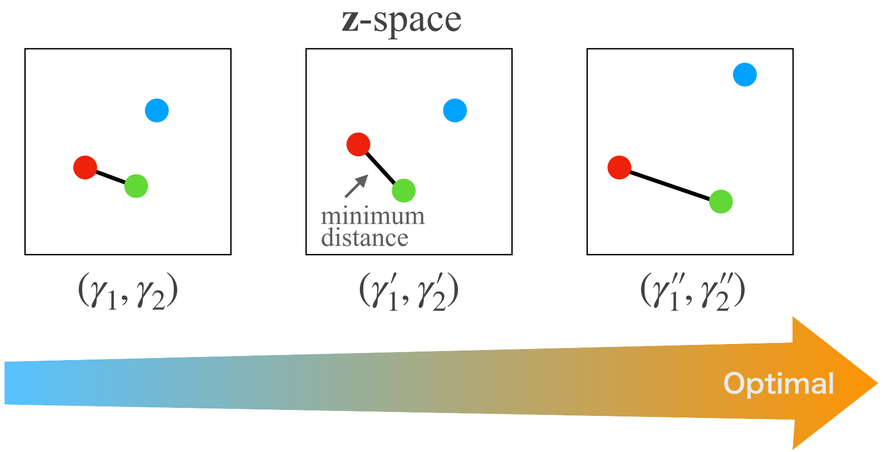
JGR paper

Latent space dimension & Lagrange multipliers:

Those paramters are determined by the following rules while ensuring the classifiability conditions:

- **The latetent space** should be the lowest dimension.
- **The optimal multipliers** are chosen from a candidate set by maximizing the minimum inter-cluster distance.



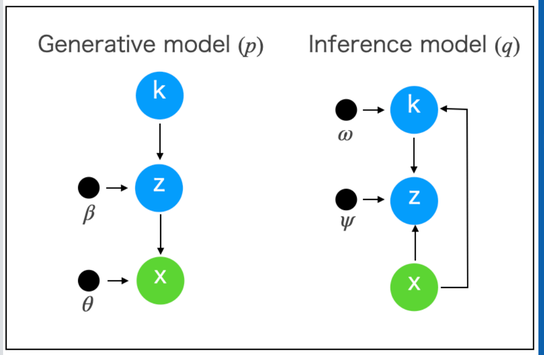


Method: Network architecture

JGR paper

Our input layer has 2-channel to handle 2-dimensional velocity vector fields.

Layer	p(z|k)	p(x|z)
Input	e_k(bs, K)	z(bs, z_dim)
Hidden	FCN, K, 1024 LeakyReLU(0.2), FCN, 1024, 2*z_dim	FCN, z_dim, 40 FCN, 40, 1024 BatchNorm1d(1024) LeakyReLU(0.2) FCN, 1024, 128 x lon//4 x lat//4 Flatten ConvTranspose2d, 128, 64, 4x4, stride=2, padding=1 LeakyReLU(0.2) ConvTranspose2d, 64, 2, 4x4, stride=2, padding=1 Linear activation
Output	mu(bs, :z_dim); Log_variance(bs, z_dim:); Reparameterize: z(mu, Log_variance)	x(bs, 2, lon, lat)



Layer	q(z|x, k)	q(k|x)
Input	x(bs, 2, lon, lat); e_k(bs, K)	x(bs, 2, lon, lat)
Hidden	BatchNorm2d, 2 Conv2d, 2, 64, 4x4, stride=2, padding=1 BatchNorm2d, 64 LeakyReLU(0.2) Conv2d, 64, 128, 4x4, stride=2, padding=1 BatchNorm2d, 128 LeakyReLU(0.2) Flatten x1: FCN, 128 x lon//4 x lat//4, 40 x2: FCN, 128 x lon//4 x lat//4, K Concat, x1, x2*k FCN, 40+K, 1024 BatchNorm1d, 1024 LeakyReLU(0.2) FCN(1024, 2*z_dim)	Conv2d, 2, 64, 4x4, stride=2, padding=1 BatchNorm2d, 64 LeakyReLU(0.2) Conv2d, 64, 128, 4x4, stride=2, padding=1 BatchNorm2d, 128 LeakyReLU(0.2) Flatten FCN, 128 x lon//4 x lat//4, 512 Dropout(prob=0.8) ReLU FCN, 512, 512 Dropout(prob=0.6) ReLU FCN, 512, K Softmax
Output	mu(bs, :z_dim); Log_variance(bs, :z_dim); Reparameterize: z(mu, Log_variance)	Gumbel Sofmax Sampling e_k(bs, K)

Generative model

Inference model

Data augmentation

JGR paper

Need of augmentation

Our sample size is only **164** (*Too small !*)

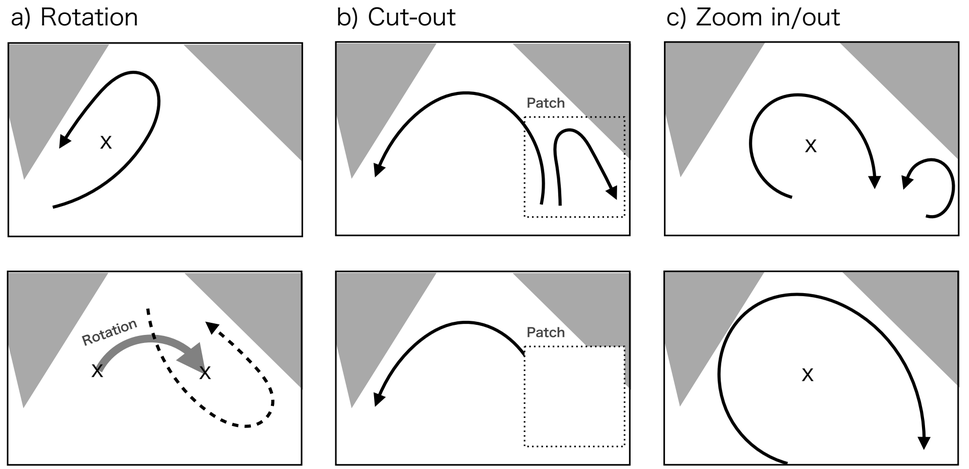
Problems in conventional methods

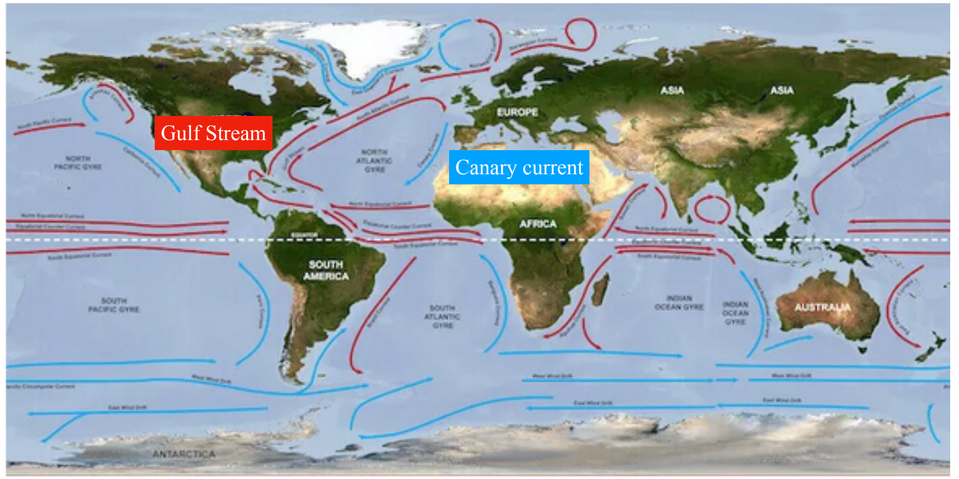
Those methods implicitly postulate an equivalence between the original and augmented data.

- Rotation, zooming: Geometrical symmetry.
- Cut-out: Robustness of data for a partial information scarcity.

However, these operations can create unrealistic pattern or miss a distinctive feature.

Rotation symmetry identifies
the Gulf Stream and Canary current
as identical..
although the underlying mechanisms
are different.





This world map shows the five oceanic gyres and how they impact ocean circulation.

Credit: NOAA

Data augmentation

JGR paper

Noise injection :

$$\mathbf{x} \to \mathbf{x} + \delta\mathbf{x} \qquad \mathbf{x} \in \mathbb{R}^d$$

PC-scaled noise injection:

$$\delta\mathbf{x} = \sum_{i=1}^{M} \varepsilon_i \sqrt{\lambda_i}\, \mathbf{v}_i$$

$\varepsilon_i \sim \mathcal{N}(0, \epsilon)$

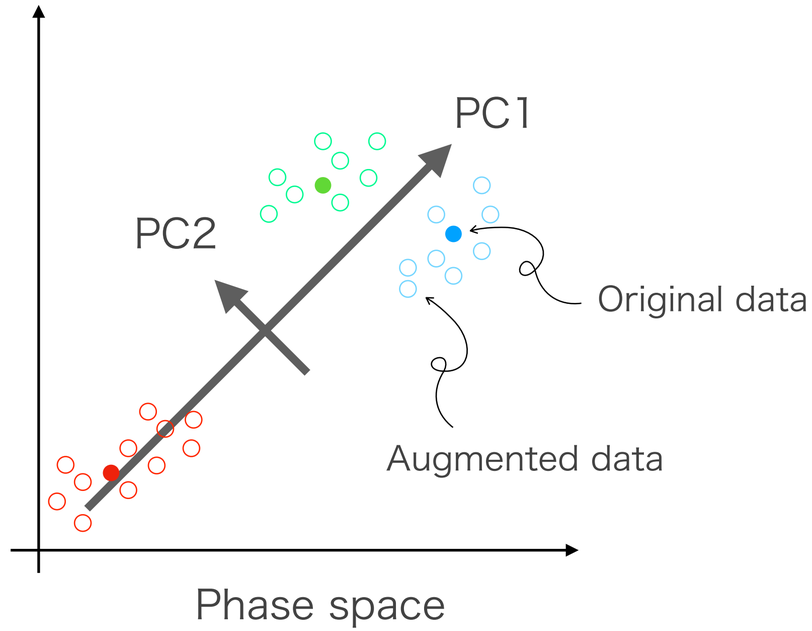
λ_i : eigenvalue

$\mathbf{v}_i \in \mathbb{R}^d$: eigenvector

$M = \max(N, d)$; N : num. of samples

i.e.,

$$\mathbf{x} \to \sum_{i=1}^{M} \left(c_i + \varepsilon_i \sqrt{\lambda_i} \right) \mathbf{v}_i \qquad c_i \text{ : coefficient}$$



Data augmentation

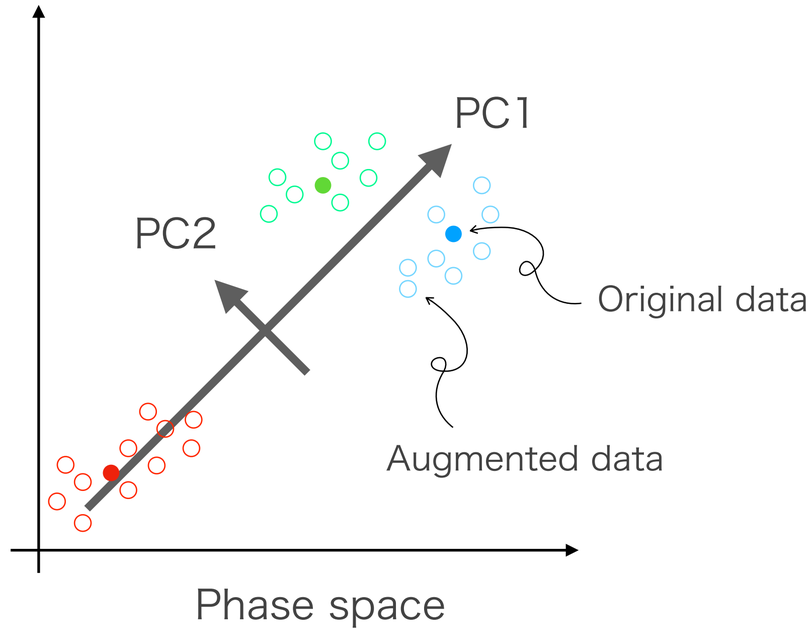
JGR paper

Noise injection :

$$\mathbf{x} \to \mathbf{x} + \delta\mathbf{x} \qquad \mathbf{x} \in \mathbb{R}^d$$

PC-scaled noise injection:

$$\delta\mathbf{x} = \sum_i^M \varepsilon_i \sqrt{\lambda_i}\mathbf{v}_i$$

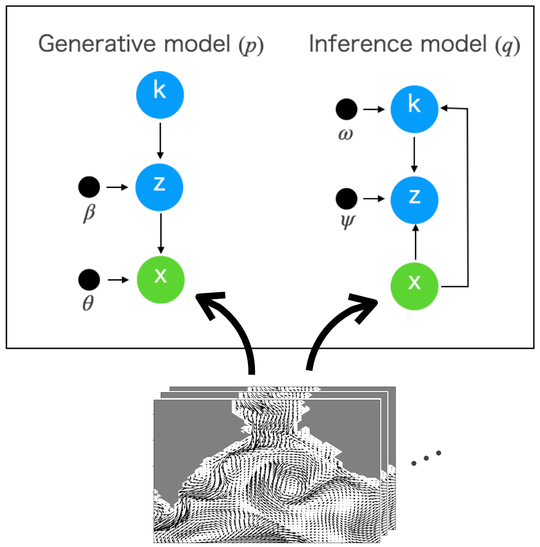


Advanteges

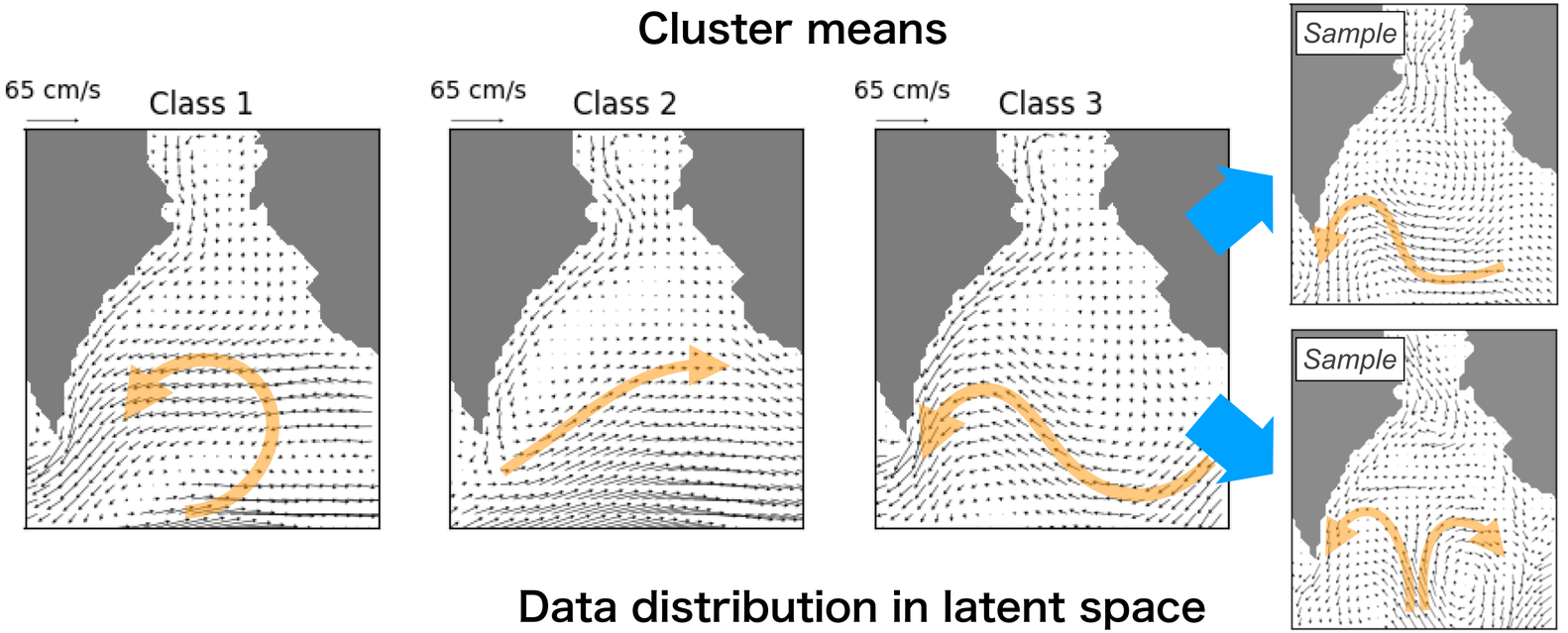
- Since M is bounded by the sample size ($N \ll d$), it is not likely that a noisy structure appears in the augmented data.
- It is likely that the information about the aggregation of data points in the phase space is preserved after the augmentation.
- It is not likely that dynamically unpreferable structures are generated by the augmentation.

Application: Clustering of Kyucho

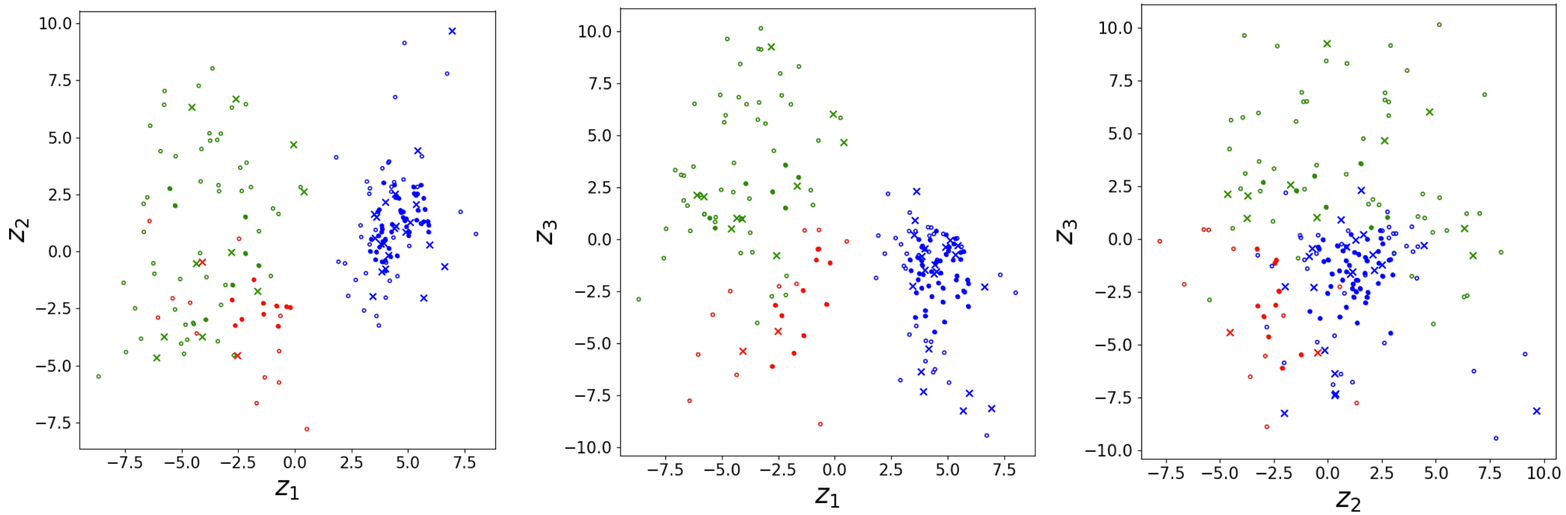
JGR paper



Cluster means

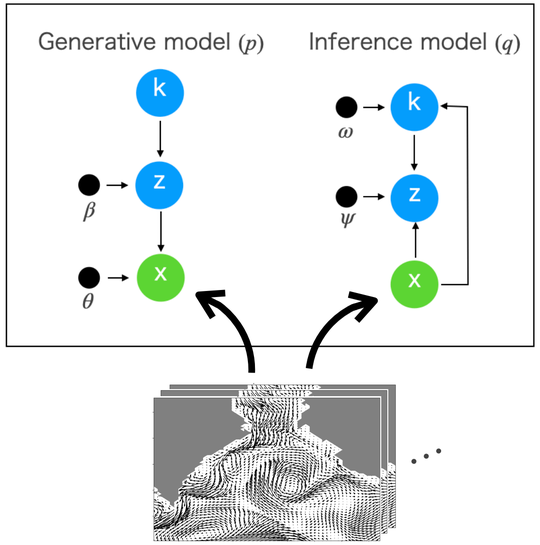


Data distribution in latent space



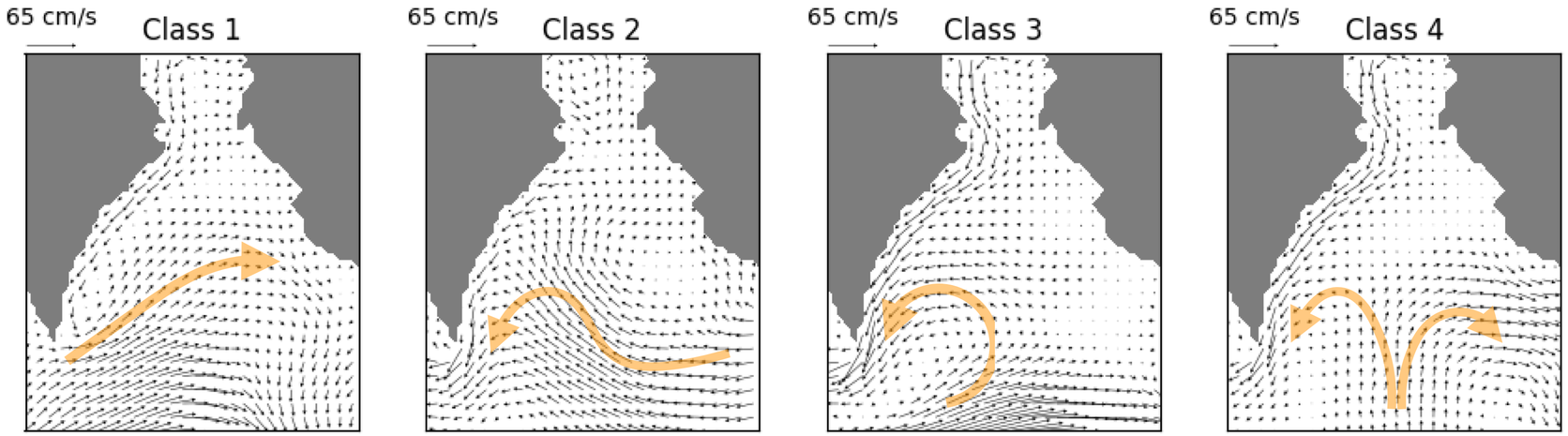
Application: Clustering of Kyucho

JGR paper

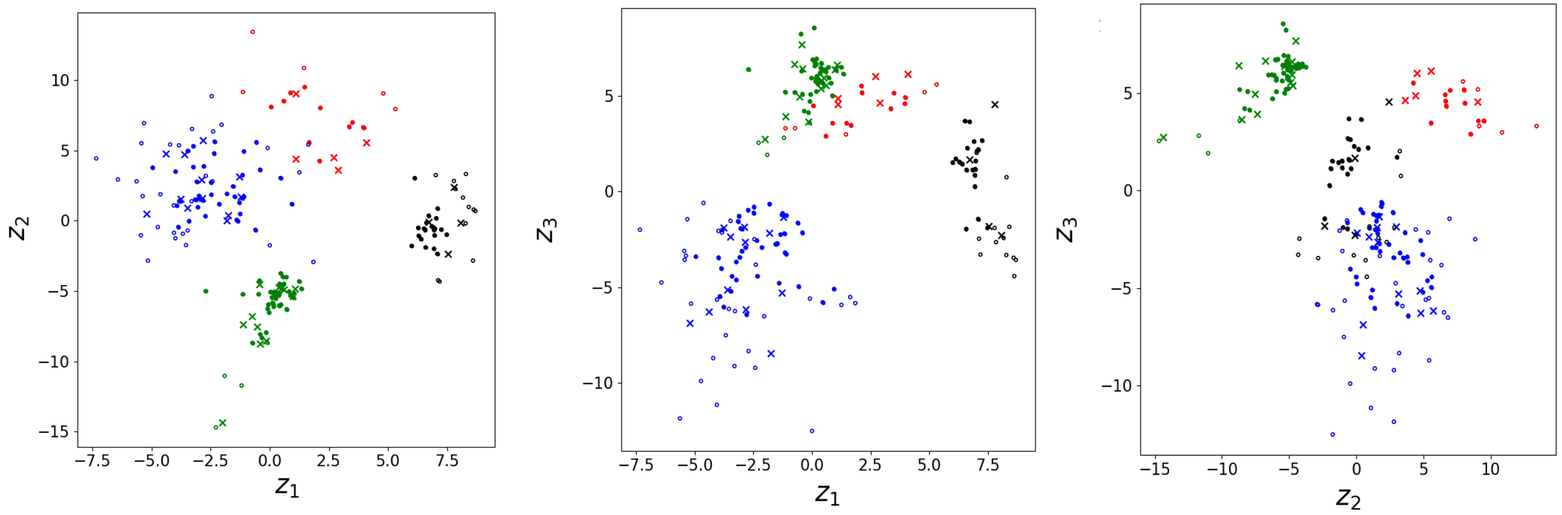


$$N = 164 + \underbrace{10 \times 164}_{\text{augmentation}}$$

Cluster means

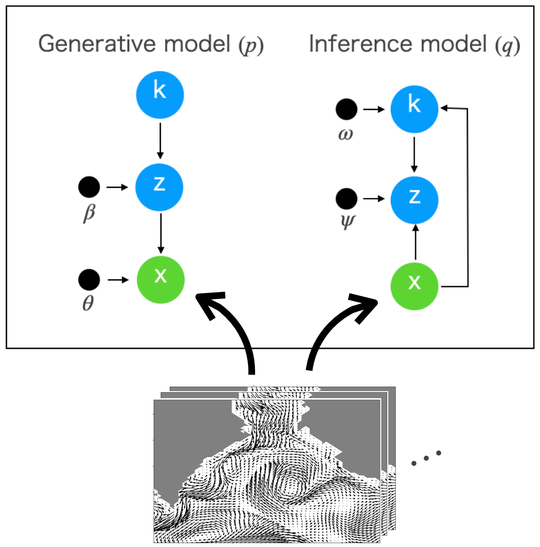


Data distribution in latent space



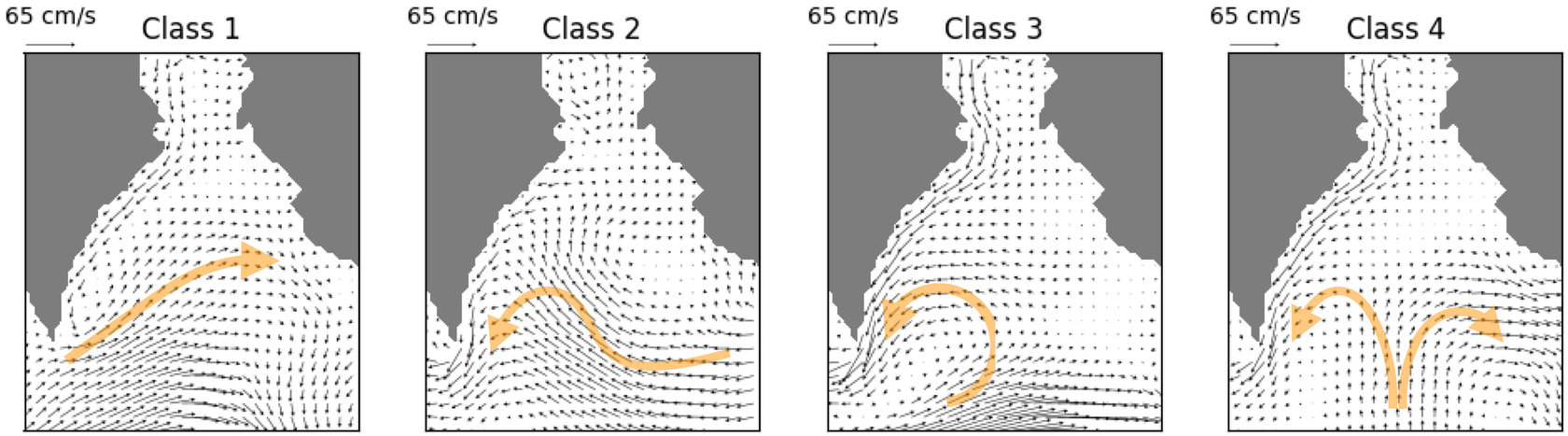
Application: Clustering of Kyucho

JGR paper

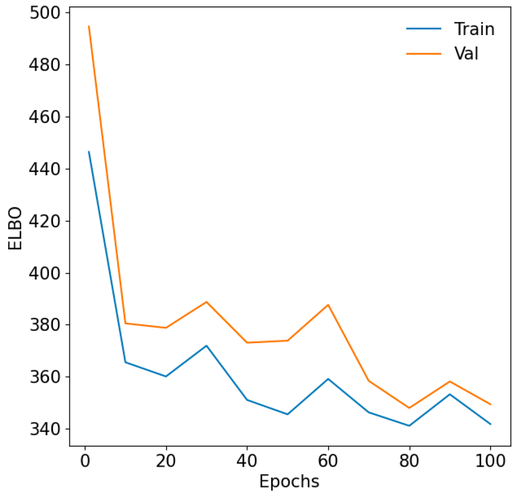


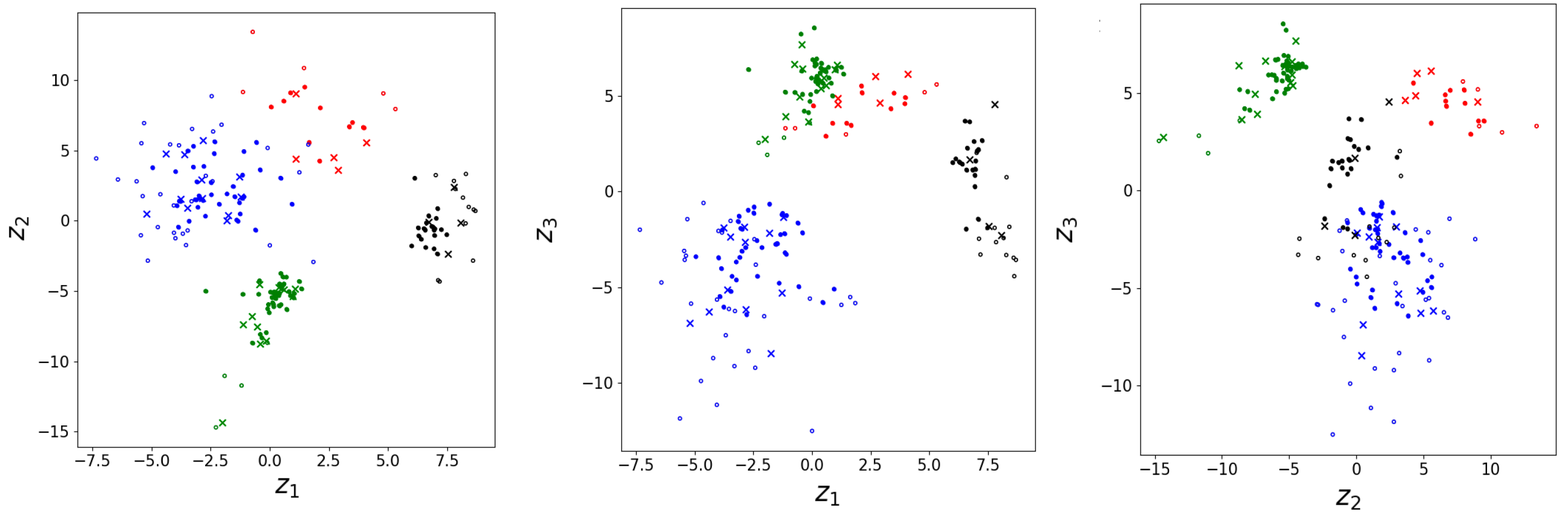
$$N = 164 + \underbrace{10 \times 164}_{\text{augmentation}}$$

Cluster means



Data distribution in latent space

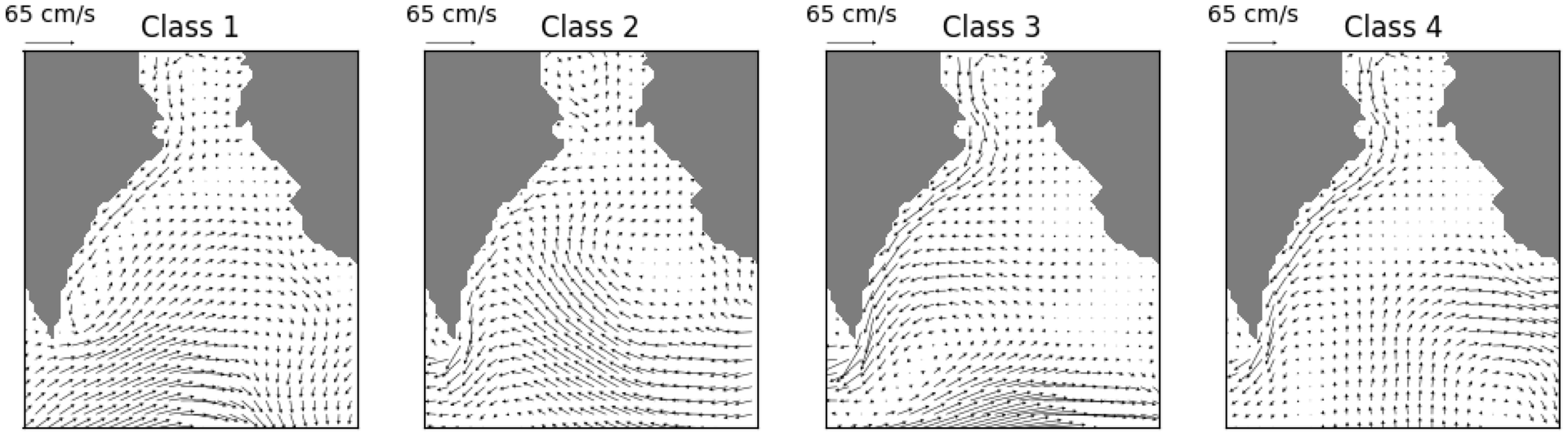




Application: Clustering of Kyucho

JGR paper

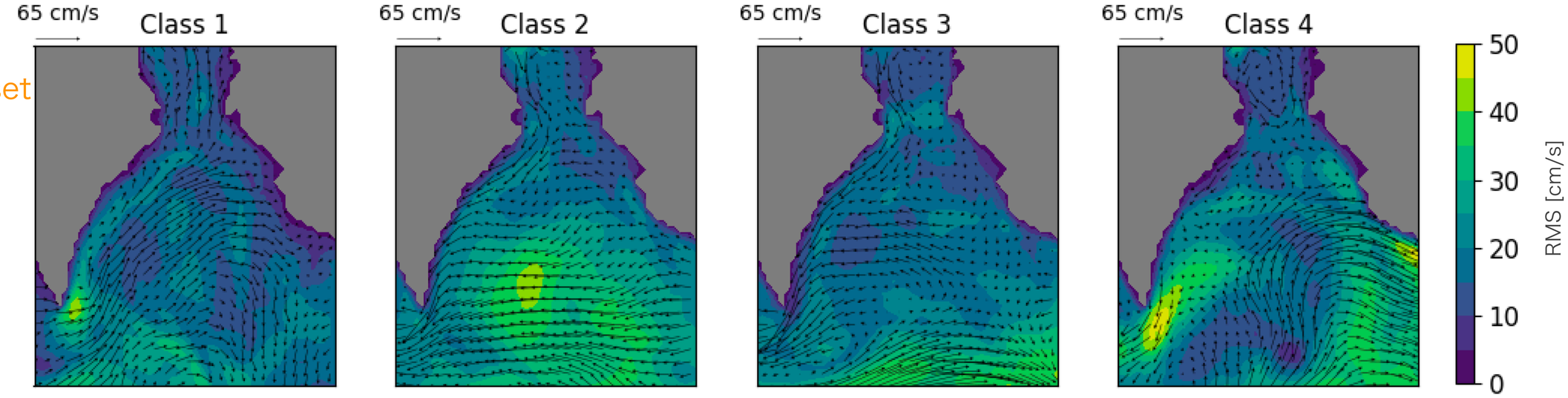
Cluster means



Reconstructions by ensemble average

Validation data set

Ensemble ave.
for each k

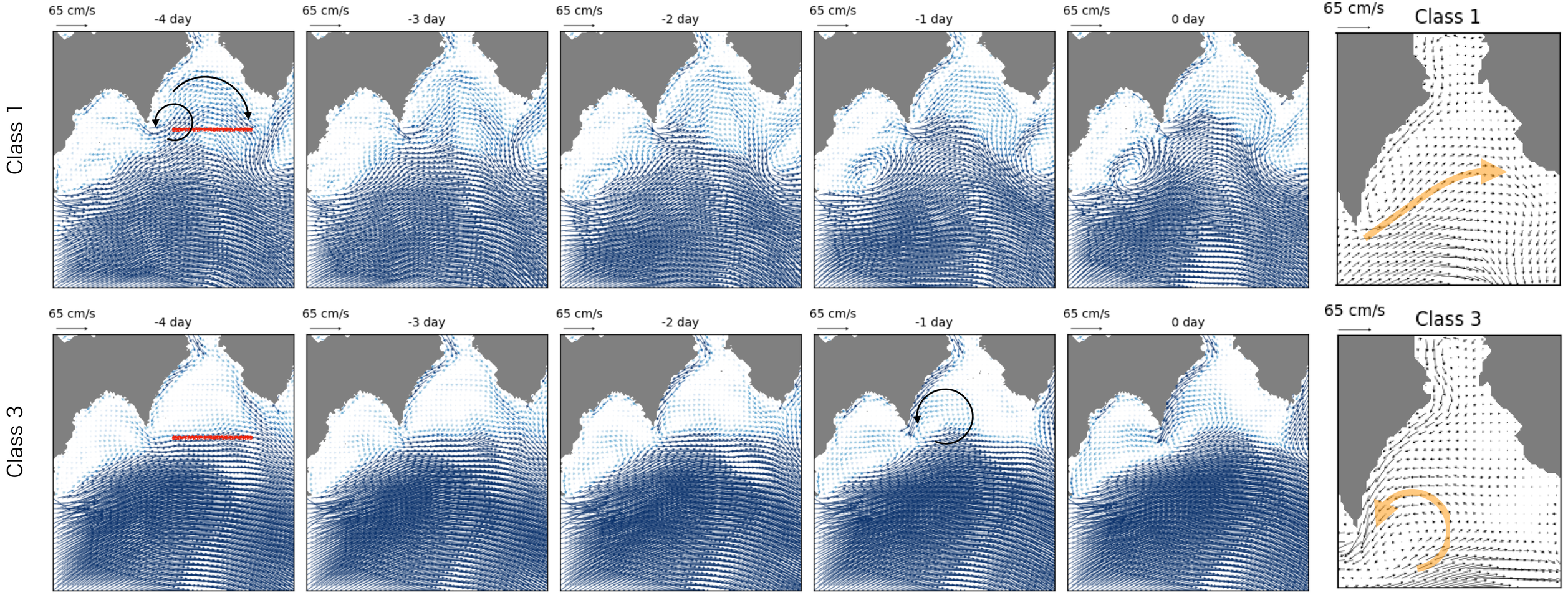


Application: Time evolutions

JGR paper

Group-wise ensemble average is performed for surface **geostrophic** velocity.

Each group consists of the members within 1σ from the cluster mean in the latent space.

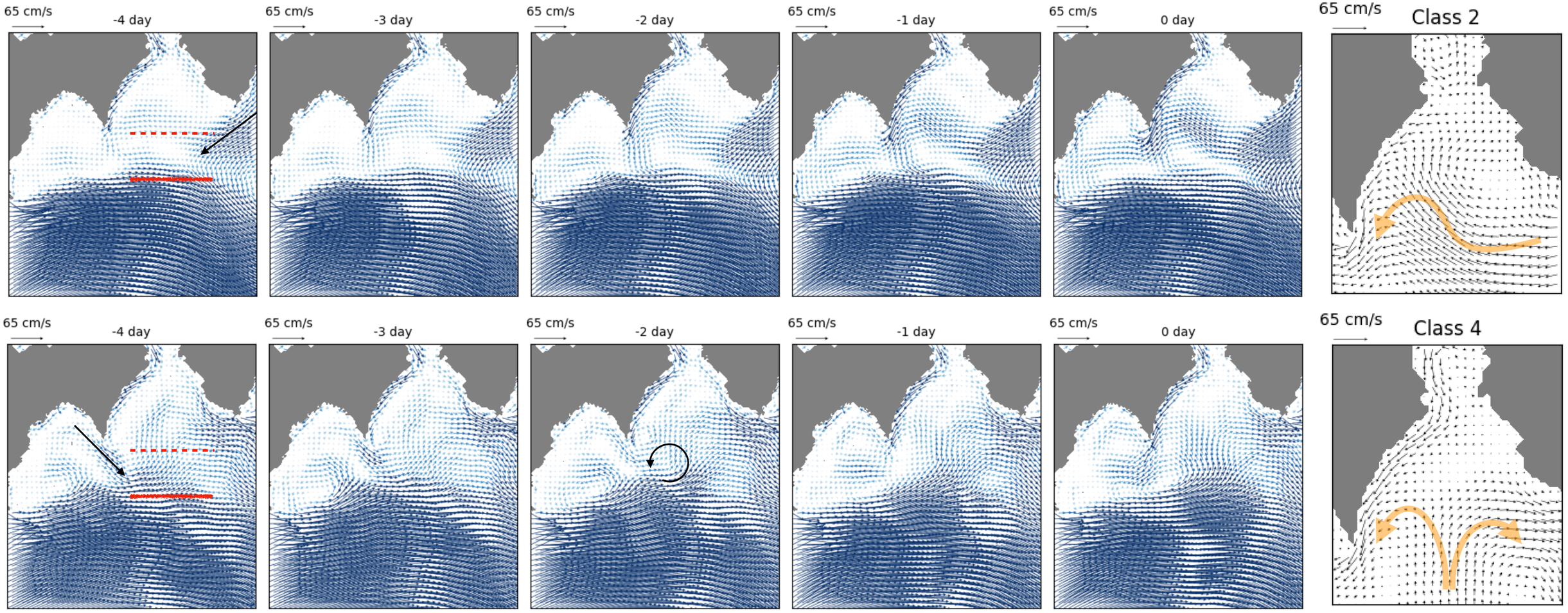


Application: Time evolutions

JGR paper

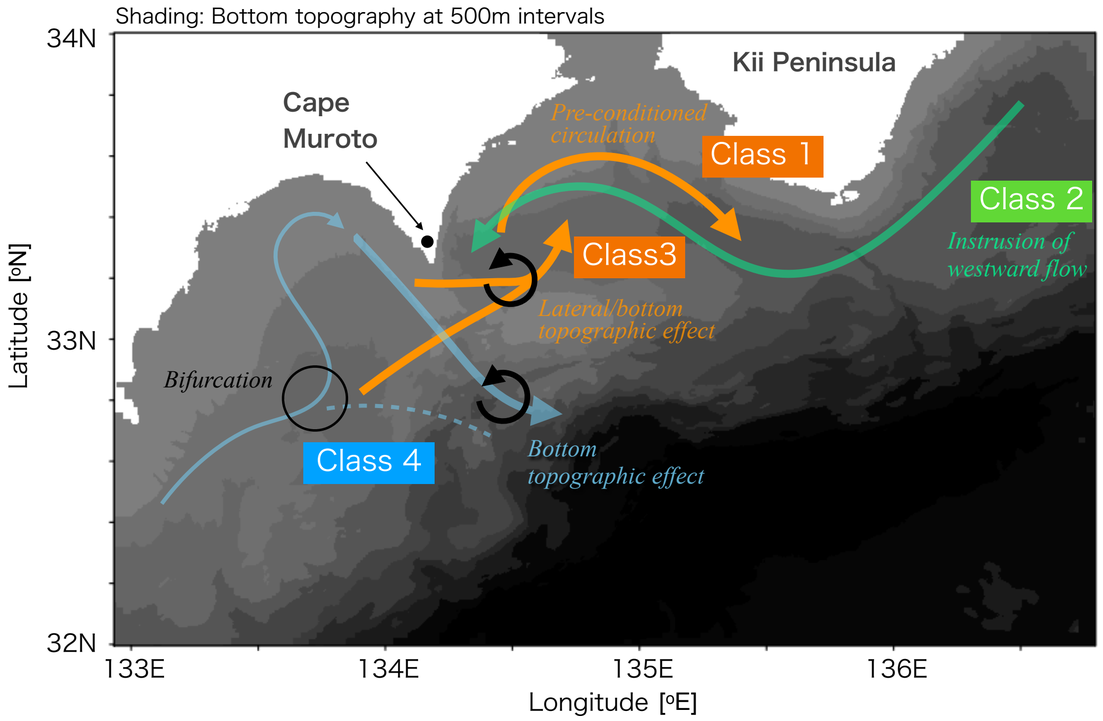
Group-wise ensemble average is performed for surface **geostrophic** velocity.

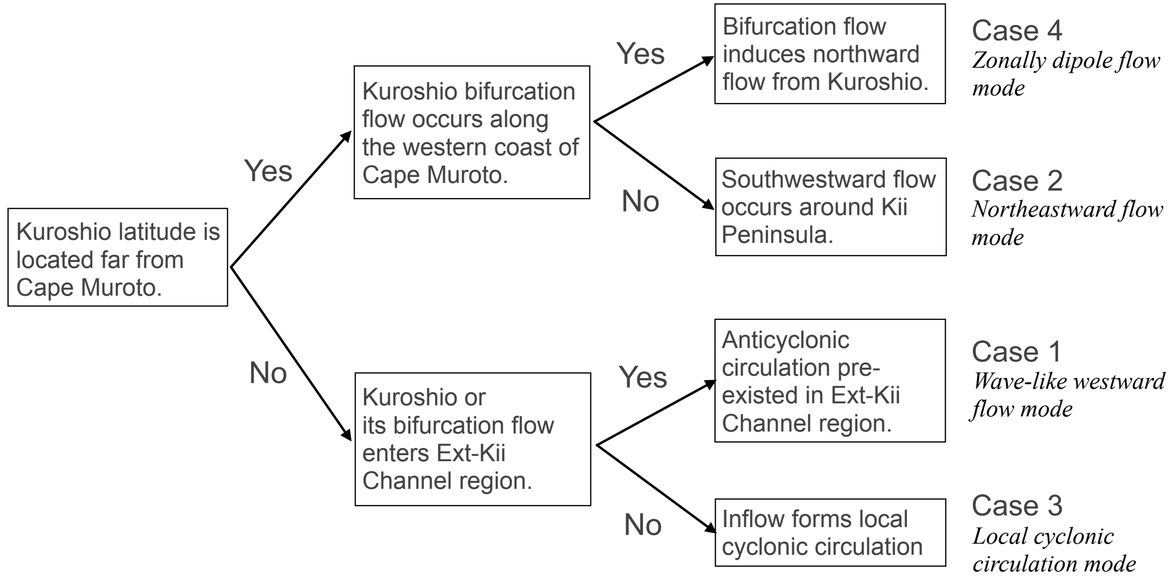
Each group consists of the members within 1σ from the cluster mean in the latent space.



Key process & potential precursors

JGR paper





Summary

JGR paper

What we have done is

This paper proposed the total method for clustering the geophysical fluid circulations on the basis of the VAE algorithm with the augmentation technique using the PC-scaled noise injection.

Clustering analysis shows

The method successfully identifies 4 distinctive modes in the circulation field accompanied by Muroto-Kyucho in spite of using the data with a small sample size.

- This is the first study to *objectively* provide the specific circulation structures for the conventional human-classified blurred flow patterns in this region.

Time-series analysis finds

Those modes are caused through different processes associated with the open ocean state variability, suggesting that those modes are dynamically discriminative.

- The identified key phenomena on those processes may provide the bases for developing a simplified model for the Kyucho prediction and discovering the detailed mechanisms essential for Kyucho.

Potential applications

- Any other geophysical fluid circulation data with a small sample size, like given in Event Attributions.
- Our 2-channel network may be used to identify non-linear coherent modes between any two variables.

Data augmentation

JGR paper

Performance test

Idealized vector field:

$$\mathbf{x} = V_1\mathbf{v}_1 + V_2\mathbf{v}_2 + V_3\mathbf{v}_3 \simeq \underline{V_1}\mathbf{v}_1 + \underline{V_2}\mathbf{v}_2$$

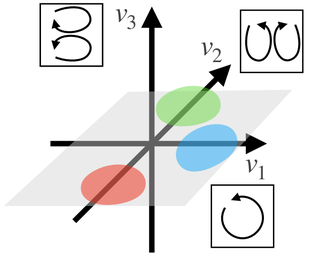
Cyclonic monopole Two-types of zonal dipole

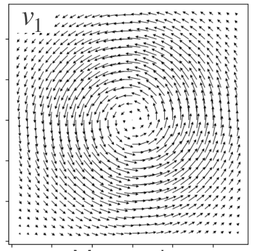
30 samples for each (90 in total)

Methods: $\mathbf{x} \to \mathbf{x} + \delta\mathbf{x}$

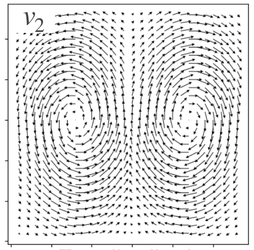
- PC-scaled $\delta\mathbf{x} = \Sigma_i \varepsilon_i \sqrt{\lambda_i}\mathbf{v}_i$
- Conventional $\delta\mathbf{x} = \alpha\sqrt{\lambda_1}$ $(\alpha \sim \mathcal{N}(0,1))$
- PC-uniform $\delta\mathbf{x} = \Sigma_i \varepsilon_i \sqrt{\lambda_1}\mathbf{v}_i$

Augmented 10-fold

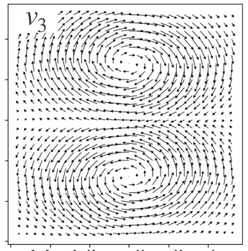




Monopole

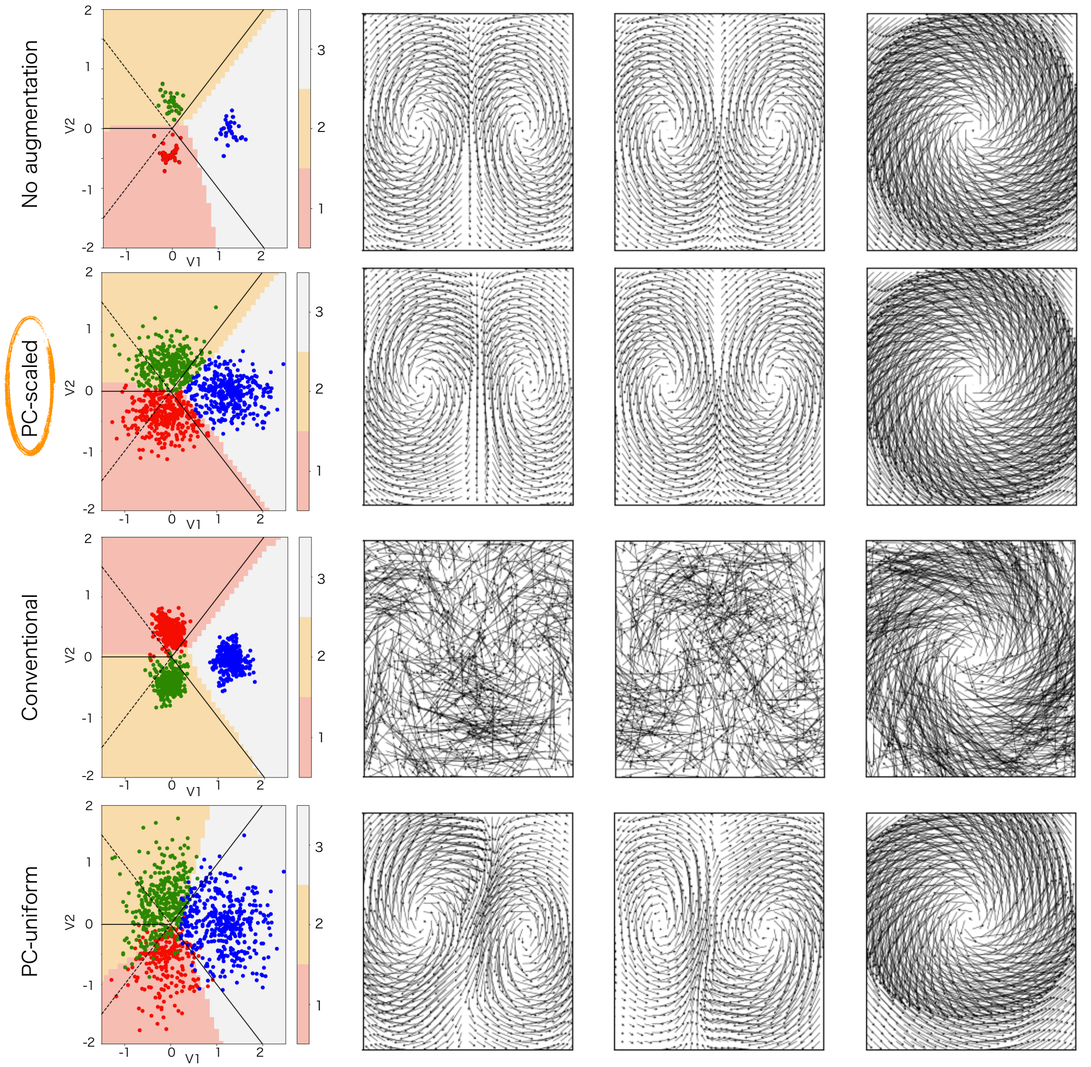


Zonally dipole



Meridionally dipole

(Add as a small noise)



(epochs=1000, minibatch size = N/3, dim(z)=2)

Data augmentation

JGR paper

Performance test

Idealized vector field:

$$\mathbf{x} = V_1\mathbf{v}_1 + V_2\mathbf{v}_2 + V_3\mathbf{v}_3 \simeq \underset{\uparrow}{V_1}\mathbf{v}_1 + \underset{\uparrow}{V_2}\mathbf{v}_2$$

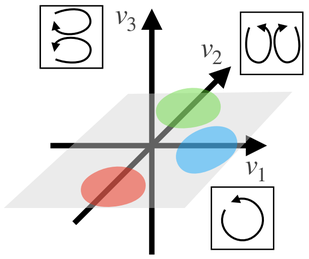
Cyclonic monopole　　Two-types of zonal dipole

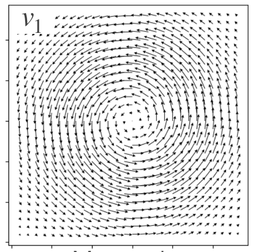
30 samples for each (90 in total)

Methods:

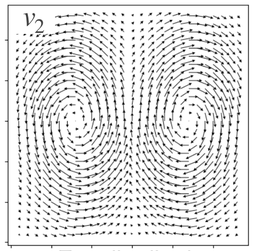
- Rotaion
- Cut-out
- Zooming

Augmented 10-fold

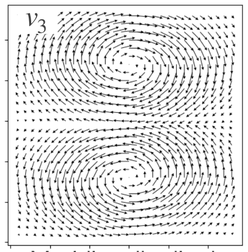




Monopole

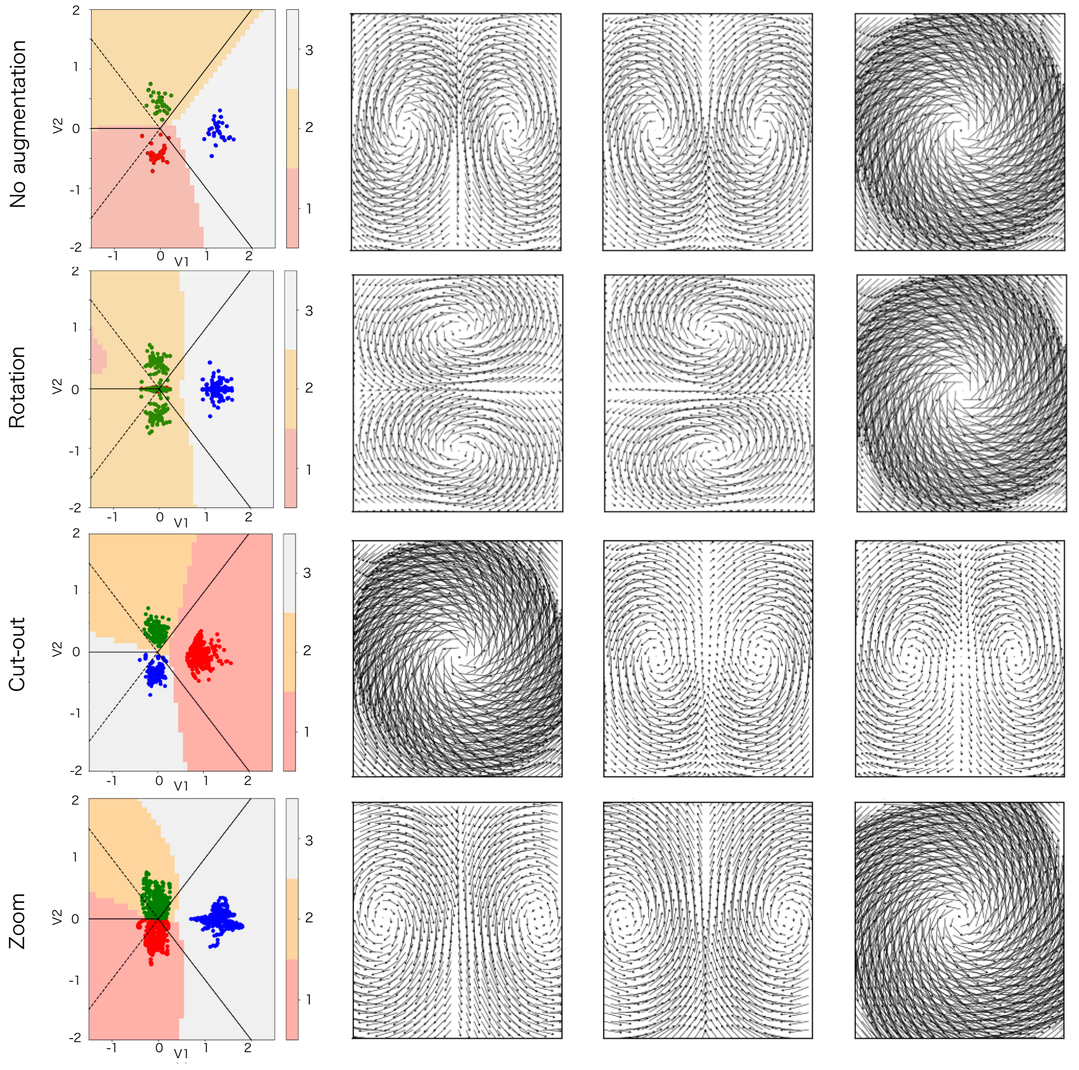


Zonally dipole



Meridionally dipole

(Add as a small noise)



(epochs=1000, minibatch size = N/3, dim(z)=2)