REPRESENTATION STABILITY IN DOMAIN SPACE: CROSS-DOMAIN VALIDATION FOR UNSUPERVISED SCIENTIFIC DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised learning is increasingly used for scientific discovery in settings where labeled data are unavailable or ill-defined. Validation is critical but challenging, and internal cluster-quality metrics encode geometric or statistical assumptions that scientific data routinely violate, and external ground truth is typically absent. This paper takes the ensemble-based NEMI workflow (Sonnewald, 2026) and develops it into a formal validation principle for unsupervised scientific discovery, evaluated across four Earth science domains.

The principle is representation stability in domain space. If a discovered structure reflects a real property of the underlying system, the partition it implies in geographic, sequence, or property space should persist under perturbations of the learned latent representation. The same ensemble that NEMI uses for clustering yields uncertainty estimates in two coordinate systems: 1) the embedding space where clustering operates, and 2) the domain space in which scientific meaning is found. We argue that the domain-space entropy field is the scientifically meaningful diagnostic, since domain-space output from any single fixed embedding is a deterministic projection of within-embedding uncertainty alone.

Across ocean ecology, ocean biogeochemistry, ocean dynamics, sea-ice, and atmospheric dynamics, low-entropy regions identify regimes that support stable scientific interpretation, while the geometry of where uncertainty concentrates generates hypotheses that can be tested. We find a recurring pattern, where low-entropy structure within regions that domain intuition identifies as transitional, suggesting candidate refinements to established regime boundaries. We also analyze sensitivity to cluster alignment strategy across domains and show that the spatial structure of the entropy field is preserved under both Greedy and Hungarian alignment, with quantifiable differences tied to split/merge events.

1 INTRODUCTION

In many scientific settings, dimensionality of data is large, the data manifold is nonlinear, and the notion of a "correct" label is ill-defined. Unsupervised learning methods are therefore widely used to explore structure in such data, often with the goal of identifying regimes, patterns, or organizing principles that can inform scientific understanding(Bishop, 2006; Goodfellow et al., 2016; Sonnewald et al., 2021).

Unlike supervised learning, unsupervised discovery lacks a canonical validation framework. Internal validation metrics quantify properties such as compactness or separation under a fixed representation, but these metrics embed assumptions that are frequently violated by scientific data (heterogeneity, anisotropy, nonstationarity, incomplete sampling)(Arbelaitz et al., 2013; Schlake & Beecks, 2024; Jenniges et al., 2025). Moreover, representation learning itself is typically necessary but uncertain, introducing an additional and often unacknowledged source of variability.

In this work, we argue that validation of unsupervised scientific discovery must explicitly account for uncertainty in the learned representation and must be evaluated where scientific meaning is expressed within its application domain, where we here use a geophysical space. Our central premise is that meaningful scientific structure should be reproducible across stochastic realizations of the latent

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

space *after projection back to domain space*, for example, geographical space as used in our case study. In general terms, the space where the interpretation should take place is the coordinate system in which scientific entities have natural relationships, constraints, or causal structure. This is typically the space where domain experts reason about the system, physical laws or biological constraints apply, and where measurements are made and predictions would be meaningful.

This problem of validation motivates an ensemble workflow that propagates representation uncertainty through clustering and exposes it explicitly as geographic uncertainty, which directly determines interpretability and downstream utility. Soft or fuzzy clustering methods, including fuzzy c-means, probabilistic spectral clustering, and Gaussian mixtures on a fixed embedding, provide per-point uncertainty without requiring an ensemble, but this uncertainty is conditional on a single learned representation, which is itself a major and largely unacknowledged source of variability in nonlinear manifold learning for scientific data. Clustering stability and consensus clustering methods (Ben-Hur et al., 2002; Lange et al., 2004; von Luxburg, 2010; Monti et al., 2003) do use ensembles, often looking at agreement using resampling or algorithmic randomness, for example in the initalization, through global stability scores or co-clustering matrices. Our workflow differs in two ways: 1) we perturb the learned representation before applying deterministic clustering, and 2) we evaluate agreement for each location in domain space, rather than as a global score. The result quantifies uncertainty over representations projected into domain space, which represents a different and complementary quantity.

The workflow extends the Native Emergent Manifold Interrogation (NEMI) framework Sonnewald (2026), with a formalized validation principle, a dual-space treatment of uncertainty, and a quantitative cross-domain analysis. Code is available at [anonymized for review]. This manuscript makes the contributions:

- **Dual-space validation as a deployed methodology.** We formalize the principle implicit in NEMI's ensemble validation as representation stability evaluated in domain space, and quantify uncertainty in both the embedding space (where clustering operates) and the domain space (where scientific meaning is defined) from a single ensemble (Fig. 1). Methods that operate on a fixed embedding can only do the former, and the domain-space output is a deterministic projection of the same quantity.

- **Scientific findings and deployment across Earth science domains.** NEMI is in active use across ocean ecology, ocean biogeochemistry, ocean dynamics, and atmospheric dynamics, with case studies appearing in domain-science. The framework is also contributing to model development toolkits aimed at improving sea-level predictability. In each domain, the entropy field has produced concrete hypotheses about regime structure that single-realization clustering does not.

- **Sensitivity analysis of alignment strategies.** Cluster correspondence across ensemble members is a non-trivial design choice that the prior NEMI work identified as future work and that has not been systematically evaluated. We compare Greedy and Hungarian alignment across five domains, showing spatial structure of the entropy field preserved under both, while absolute entropy values differ in ways tied to split/merge events.

2 Workflow Overview

The NEMI workflow comprises four stages: 1) stochastic embedding, 2) deterministic clustering in latent space, 3) ensemble construction, and 4) projection back to domain space (Fig. 1). The premise is that meaningful scientific structure should be reproducible across stochastic realizations of the latent representation, and that regimes that reflect *real* physical phenomena should correspond to coherent regions in domain space and *persist* across plausible embeddings, while a regime visible only in one particular embedding is likely a representation artifact. Throughout, embedding space (where clustering operates) is treated as distinct from domain space (where scientific meaning is defined), and robustness is evaluated in the latter, since regions, boundaries, and overlap are scientifically meaningful only there.

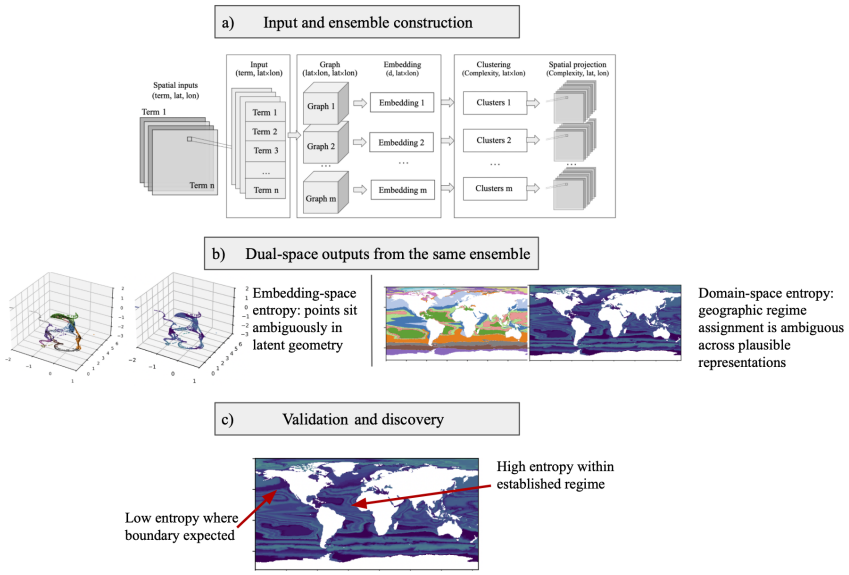


Figure 1: **The NEMI framework for dual-space validation and discovery.** a) Spatial inputs pass through m stochastic embeddings, each clustered and projected back to physical coordinates, b) The same ensemble yields entropy fields in two coordinate systems: embedding space (where points are ambiguous in latent geometry) and domain space (where regime assignment is ambiguous across plausible representations), c) The domain-space entropy H_i validates regime robustness (low H_i, left) and enables a user to discover signals where uncertainty concentrates unexpectedly (right). a) is modified from Sonnewald (2026).

3 METHOD

NEMI was introduced in Sonnewald (2026) as a workflow combining stochastic manifold learning, deterministic clustering, and ensemble-based uncertainty quantification, with informal validation through visual inspection, ensemble agreement, and domain-specific expectation. We retain the workflow but reinterpret the role of ensemble variability and formalize the validation principle it instantiates. Where Sonnewald (2026) frames stochasticity as a regularizer smoothing the optimization landscape, we treat it as a sampling mechanism over plausible latent representations, and use the resulting domain-space ensemble disagreement as a validation diagnostic in its own right.

3.1 STOCHASTIC MANIFOLD EMBEDDING

Let $\{\mathbf{x}_i\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^D$, denote observations of a complex physical system, modeled as samples on or near a low-dimensional manifold $\mathcal{M} \subset \mathbb{R}^D$. A stochastic manifold learning algorithm defines a mapping

$$\Phi_\theta : \mathbb{R}^D \to \mathbb{R}^d, \tag{1}$$

where θ denotes stochastic elements (random initialization, sampling during optimization). Each realization yields an embedding

$$\mathbf{z}_i^{(k)} = \Phi_{\theta_k}(\mathbf{x}_i), \quad k = 1, \ldots, K. \tag{2}$$

Rather than treating any single embedding as definitive, we treat $\{\mathbf{z}_i^{(k)}\}$ as samples from a distribution over plausible latent embeddings consistent with the data.

We instantiate Φ_θ with UMAP (McInnes et al., 2018), following (Sonnewald, 2026). UMAP's stochasticity (negative sampling, random initialization) does not correspond to a Bayesian posterior. We rely instead on the operational property that nonlinear manifold learning on scientific data is non-identifiable: many distinct low-dimensional embeddings are consistent with the same local neighborhood structure, and the algorithm has no principled basis for selecting among them. Algorithmic perturbations traverse this equivalence class, and the ensemble samples plausible embeddings

rather than recovering a posterior. What matters for the validation principle is the consequences of this variability in domain space, which is that structures that persist across plausible embeddings are robust, and that those that do not persist are not. The framework is agnostic to the source of stochasticity, and any probabilistic latent variable model with an explicit posterior may be substituted for UMAP.

Hyperparameters (n_neighbors and min_dist) are selected using the domain-space entropy as a guide and then held fixed within each ensemble, so that ensemble variability arises only from stochastic UMAP initialization and optimization. Because the same diagnostic informs tuning and interpretation, we treat low entropy as necessary but not sufficient evidence of robust structure.

3.2 Clustering in Latent Space

For each embedding realization, we apply a deterministic clustering operator

$$\mathbf{c}^{(k)} = \mathcal{C}(\{\mathbf{z}_i^{(k)}\}), \quad \mathbf{c}^{(k)} := \{c_i^{(k)}\}_{i=1}^{N}, \tag{3}$$

where $\mathcal{C}$ maps embedded points to discrete labels $c_i^{(k)} \in \{1, \ldots, C\}$. Determinism in $\mathcal{C}$ isolates the uncertainty arising from representation learning rather than the clustering procedure. We use hierarchical agglomerative clustering for the ocean dynamics, ecology, and atmospheric dynamic case study and DBSCAN for the biochemistry case study. The framework does not depend on the choice of clustering method.

3.3 Projection to Domain Space and Ensemble Construction

The index i corresponds to a physical location (e.g., a grid cell at longitude–latitude (θ_i, ϕ_i)) and time, if applicable. Each ensemble member k induces a categorical label field over the domain by assigning the latent-space cluster label $c_i^{(k)}$ to the corresponding physical location. Repeating this procedure yields an ensemble of label fields $\{\mathbf{c}^{(k)}\}_{k=1}^{K}$, where each member is a hypothesis about the partition implied by a plausible latent representation. The ensemble encodes epistemic uncertainty associated with representation learning, expressed as uncertainty over geographic label fields.

3.4 Label Alignment and Geographic Consensus

Cluster labels are defined only up to permutation across ensemble members, so ensemble statistics in domain space require establishing correspondence between clusters across runs. In global geophysical applications this is inherently a *global* matching problem, and correspondence must be determined using full spatial label fields rather than local proximity in the embedding, because embedding coordinates vary across realizations and carry no consistent meaning across runs. Matching is also non-trivial when regimes split or merge between realizations, in which case no one-to-one correspondence exists and any alignment imposes a choice. The prior NEMI work (Sonnewald, 2026) identified embedding alignment as future work without specifying or evaluating a particular strategy, and we formalize and quantitatively compare two natural choices.

Greedy alignment. Given a reference partition $\mathbf{c}^{(0)}$ (typically the first ensemble member) and a target partition $\mathbf{c}^{(k)}$, the greedy procedure iteratively selects the pair of clusters (a, b) with the largest spatial overlap

$$(a^*, b^*) = \arg \max_{(a,b) \in A \times B} \sum_{i=1}^{N} \mathbb{I}[c_i^{(0)} = a]\, \mathbb{I}[c_i^{(k)} = b], \tag{4}$$

relabels b^* to a^*, removes both from the active sets A and B, and repeats until one set is empty. Unmatched clusters retain new labels rather than being forced into an existing assignment. The procedure is fast ($O(C^2)$ per ensemble member) and tolerates regimes that split or merge between realizations, at the cost of being globally suboptimal under a one-to-one constraint.

Hungarian alignment. The Hungarian (Kuhn–Munkres) procedure (Kuhn, 1955) solves the assignment problem exactly by finding a bijection $\pi : \{1, \ldots, C\} \to \{1, \ldots, C\}$ that maximizes total spatial overlap:

$$\pi^* = \arg \max_{\pi} \sum_{a=1}^{C} \sum_{i=1}^{N} \mathbb{I}[c_i^{(0)} = a]\, \mathbb{I}[c_i^{(k)} = \pi(a)]. \tag{5}$$

The procedure is globally optimal under the one-to-one constraint and runs in $O(C^3)$. The bijection requirement, however, is inappropriate when the effective number of regimes differs across realizations, since it forces every cluster in one partition into a unique counterpart in the other even when no genuine match exists.

Empirical entropy and consensus. After alignment, we define for each location i the empirical cluster membership distribution

$$p_i(c) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}[c_i^{(k)} = c], \quad c \in \{1, \ldots, C\}, \tag{6}$$

the consensus label $\hat{c}_i = \arg\max_c p_i(c)$, and the geographic entropy

$$H_i = -\sum_{c=1}^{C} p_i(c) \log p_i(c). \tag{7}$$

Low H_i indicates a location consistently assigned to the same regime across the ensemble, while high H_i indicates ambiguous or mixed membership. We report entropy normalized by its maximum value,

$$\tilde{H}_i = 100 \times \frac{H_i}{\log C} \quad [\%], \tag{8}$$

so that $\tilde{H}_i = 0\%$ when all ensemble members assign location i to the same regime and $\tilde{H}_i = 100\%$ when assignments are uniformly distributed across all C regimes. This allows entropy to be comparable across partitions with different numbers of clusters. The entropy field treats regime overlap as ensemble disagreement in domain space, which is the quantity that determines downstream scientific utility.

Why entropy?

Alternative overlap and similarity measures including pairwise overlap (Manning, 2008), Adjusted Rand Index (Hubert & Arabie, 1985), and Normalized Mutual Information (Strehl & Ghosh, 2002) were systematically evaluated for biogeochemical data in Jenniges et al. (2025), who found that aggregate similarity metrics can mask substantial cell-level uncertainty (e.g., overlap of 88.8% co-occurring with cell-wise uncertainty of 15.5±20%), and that internal cluster validity indices systematically disagree both with each other and with external visual validation. Entropy H_i avoids the aggregation problem by quantifying ambiguity per location, and avoids the CVI assumption-bias problem by not encoding a notion of cluster shape, density, or neighborhood structure.

Sensitivity of H_i to alignment. We evaluate sensitivity empirically across ocean ecology, ocean dynamics, ocean biogeochemistry, and atmospheric dynamics (Section 5, Table 1, Fig. 2). The spatial structure of H_i is preserved under both strategies, with high-entropy regions remaining high-entropy and low-entropy regions remaining low-entropy. Hungarian alignment yields lower mean entropy than Greedy in four of five settings, consistent with its global optimality under the one-to-one constraint, where maximizing total spatial overlap directly increases the consistency of label assignment across runs, which can therefore reduce per-location entropy. The exception is ocean ecology, where the bijection constraint forces two clusters with no genuine counterpart into spurious matches that collapse to near-zero entropy (Appendix A). If excluding the degenerate cases, Greedy yields lower mean entropy in the remaining clusters. The two strategies therefore occupy distinct regimes, where Hungarian is preferable when the effective number of regimes is stable across realizations, such as when they are set using agglomerative clustering, and Greedy is preferable when regimes are expected to split or merge. The spatial entropy field is a stable diagnostic regardless of which strategy is chosen, and the alignment-strategy comparison itself is informative about the cluster-stability regime of a given application.

4 Validation in Domain Space via Representation Stability

The central validation principle in this paper is representation stability evaluated in domain space. Clusters are computed in latent space, but validation and scientific interpretation occur after projection into geography (Fig. 1). Concretely, we assess whether the *geographic* regions implied by clustering

Domain	Clusters	Mean Entropy (%)		ΔMean
		Hungarian	Greedy	
Ocean ecology[†]	10	28.81 ± 9.19	25.55 ± 9.99	-3.26
Ocean dynamics	6	32.07 ± 19.37	35.88 ± 22.02	3.81
Ocean dynamics	15	28.24 ± 17.53	29.25 ± 16.98	1.01
Ocean biogeochemistry[‡]	310	8.53 ± 10.59	19.75 ± 17.14	11.22
Atmospheric dynamics	8	23.39 ± 2.36	26.85 ± 4.65	3.46

[†]Two degenerate clusters excluded; Hungarian alignment collapsed these to near-zero
[‡]The larger ΔMean reflects alignment difficulty for numerous small clusters ($N \leq 5$, 12% of clusters, <1% of grid cells). The spatial entropy field is largely unaffected (Fig. 2).

Table 1: **Alignment strategy comparison across domains.** Mean normalized entropy (%) computed over identical ensembles using greedy vs. Hungarian alignment. Spatial entropy patterns are consistent across methods despite differences in absolute values, confirming that the entropy field reflects data structure rather than alignment artifacts.

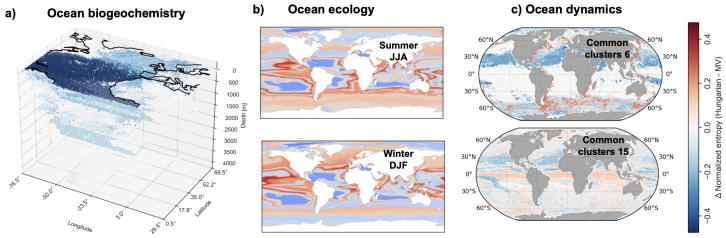


Figure 2: **Alignment strategy sensitivity across domains.** a) Ocean biogeochemistry data shown in physical coordinates, b) Ocean ecology cluster assignments, c) Ocean dynamics consensus partitions at 6 and 15 clusters, with the per-location difference in normalized entropy between alignment strategies (Hungarian - Greedy). Negative/positive values (blue/red) indicate locations where Hungarian assigns lower/higher entropy than Greedy. Note consistency with Table 1.

persist under perturbations of the latent representation. The entropy field H_i is therefore a validation diagnostic, as it reveals where the discovered structure is robust (low H_i) and where it is ambiguous or overlapping (high H_i) in the domain where scientific utility is defined.

This perspective differs from standard internal validation indices, which assess cluster geometry under a fixed representation. The unreliability of such indices for unsupervised scientific discovery has been demonstrated empirically in Jenniges et al. (2025), who showed that classical CVIs (Calinski–Harabasz, Davies–Bouldin, Silhouette) and structural CVIs (k-DBCV, CVNNH, CDR) systematically disagree on clustering quality and reflect their own embedded assumptions about cluster shape rather than data structure. Representation stability evaluated in domain space sidesteps this entirely: it does not require the analyst to commit to any particular notion of cluster geometry. Robustness is instead evaluated by the reproducibility of *projected geographic structure* under stochastic perturbations of the learned latent space, which is a distinction that matters because regions, boundaries, and overlap are meaningful only in domain space.

The geometry of domain-space uncertainty as scientific signal. Uncertainty does not vanish under projection from embedding to domain space; it reorganizes, and the resulting geometry is informative (Fig. 1c). A diffuse region of embedding-space disagreement can project into a sharply localized streak in domain space, into scattered points along a known boundary, or into coherent low-entropy structure where domain intuition expected ambiguity. The last case is particularly informative: when a region that domain experts identify as a transitional zone shows consistently low entropy across the ensemble, the ensemble is reporting that the data support a coherent regime assignment there, even if the scientific community has historically treated it as a boundary. We interpret such cases not as artifacts of the framework but as candidate sites for refinement of domain understanding. The validation principle therefore generates hypotheses, not just confidence intervals.

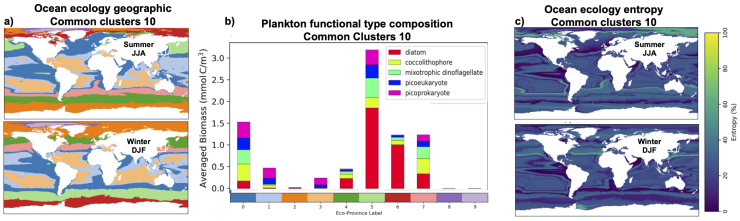


Figure 3: **Ocean Plankton Ecology.** Ecological regions a) in combined winter (DJF) and summer (JJA). The ecological make-up of regions is given by the biomass of types of plankton present b), and the associated entropy in c).

5 Cross-domain results: Geographic Utility and Diagnostics

We apply NEMI to five geophysical datasets characterized by high dimensionality, nonlinear interactions, and spatial heterogeneity. Each demonstrates a distinct type of scientific insight from the domain-space entropy field: distinguishing robust from contested boundaries (ocean ecology), guiding the choice of partitioning complexity (ocean dynamics), identifying the partition complexity at which known physical structures emerge (atmospheric dynamics), and locating informative gaps in observational data (ocean biogeochemistry).

Ocean Ecology: Distinguishing Robust from Contested Boundaries. This application identifies eco-provinces in ocean plankton, which are microscopic organisms at the base of the marine food chain, whose distributions are spatially inhomogeneous and seasonally variable (Fig. 3). Different functional types play distinct ecological roles: picoeukaryotes make nitrogen available in oligotrophic regions such as the subtropical gyres (eco-province 3, expected to remain seasonally stable), while diatoms in eco-province 5 vary strongly between boreal summer and winter and support higher trophic levels including fisheries (Sonnewald et al., 2020).

Insight. Several regions that domain intuition would identify as ecological boundaries, narrow zones between known eco-provinces, frontal regions, and transitional habitat, consistently show low entropy across the ensemble. The "stripes" of low entropy in Fig. 3c fall within the eco-provinces in panel a: the ensemble robustly assigns these streaks to a single regime regardless of which plausible embedding is realized. We interpret this not as a failure of the framework to detect known boundaries but as candidate refinements to the boundary structure itself as new scientific insight. Regions of high entropy occasionally appear in the interiors of established eco-provinces, indicating internal heterogeneity that single-realization clustering would obscure. The findings reported here form the basis of a manuscript currently under peer review.

Ocean Dynamics: Entropy as a Guide to Complexity. Ocean dynamical regimes, coherent regions where similar physical mechanisms dominate the movement of the ocean, have been an organizing concept in physical oceanography for decades (Sonnewald et al., 2019; 2023). The focus here is on a methodological question the prior work raised but did not formalize: Can the domain-space entropy field guide the choice of partitioning complexity?

Insight. A common practical question in unsupervised discovery is how many clusters to request. The entropy field provides a spatially resolved answer comparing increasing the number of clusters from 6 to 15 reduces entropy globally (Fig. 4), indicating that finer partitioning better captures the underlying regime structure. More usefully, the entropy field reveals that the appropriate complexity level varies by region. The Antarctic margin is one example: bottom-topography-driven dynamics produce low-entropy structure at fine partitioning levels in regions of small areal extent but considerable scientific importance, in contrast to the broad subtropical gyres which are stable under coarse partitioning. This framing of dynamical regimes has previously led to scientific insight (Sonnewald et al., 2019; 2023).

Atmospheric Dynamics: Emergence of Physically Meaningful Structure. The atmospheric case study uses European Centre for Medium-Range Weather Forecasts reanalysis v5 (ERA5) (Hersbach et al., 2018), a gridded product where a numerical model is constrained by observations to provide all physical fields. The features are terms in the atmospheric momentum balance, analyzed at distinct complexity levels (Fig. 5).

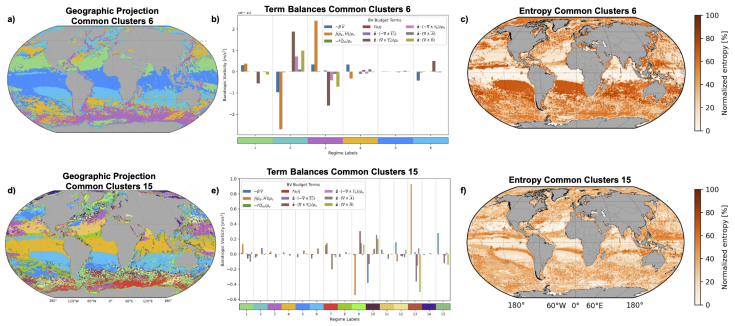


Figure 4: **Ocean dynamics.** Although clusters are computed in latent space, their scientific meaning emerges only after projection into physical space (a and d). Entropy H_i (c and f) quantifies spatial overlap and ambiguity between regimes. Cluster-wise diagnostic balances connect discovered structure to physically interpretable mechanisms (b and e).

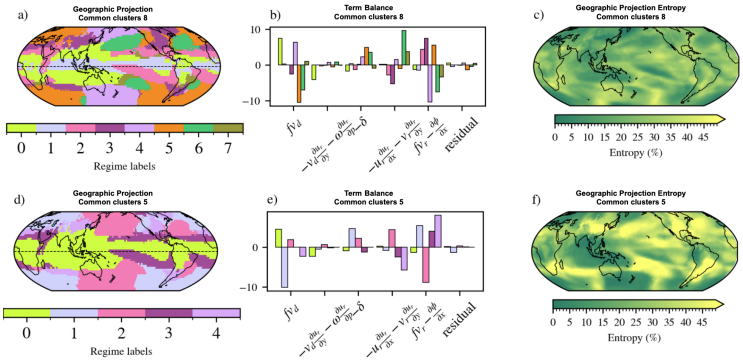


Figure 5: **Atmospheric dynamics**: ERA5 momentum-balance data analyzed at two complexity levels: a) consensus partition with 8 clusters, b) the corresponding term-balance signature for each cluster, with c) the geographic entropy field H_i. d–f) The same analysis at 5 clusters. Note, increasing clusters raises global mean, but allows known atmospheric dynamics to emerge.

Insight. Unlike the ocean dynamics case, increasing the number of clusters here *increases* mean entropy, indicating that the atmospheric system supports fewer distinct regimes before ensemble assignments become less stable. The directional difference between the two cases is itself informative, as the entropy–cluster-count relationship reflects the data, not the framework. At a sufficient complexity level, coherent structures corresponding to known dynamical features become identifiable, including equatorial circulation associated with monsoon dynamics (Fig. 5a). The framework therefore identifies the complexity level at which specific physical phenomena become distinguishable, while the entropy field reports on the stability of the resulting partition (McDevitt et al., 2026).

Ocean Biogeochemistry: Identifying Informative Data Gaps. The data come from the COMFORT effort (Korablev et al., 2022), synthesizing observations from ships and autonomous platforms over 0–5,000 m depth, with features comprising temperature, salinity, and concentrations of oxygen, nitrate, silicate, and phosphate. Unlike the other case studies, observational data here is sparse and expensive to collect: existing measurements are scientifically valuable, and the question is where additional sampling would be most informative. The UMAP-DBSCAN pipeline used here was established for these data through systematic clustering-method comparison in Jenniges et al. (2025), who reported a mean cell-wise uncertainty of $15.5 \pm 20\%$ across 100 ensemble members.

Insight. Figure 6 shows the ensemble result in three coordinate systems: the embedding (panels a, b), the geographic projection (panel c), and the geographic entropy field (panel d). High entropy concentrates at the boundary between the surface mixed layer and deeper water masses (visible as a thin band in panel c). In the embedding, the same region appears as a sparsely sampled boundary between two well-populated structures. The ocean rarely has sharp property boundaries, and observational evidence suggests this transition zone should be smoother than, for example,

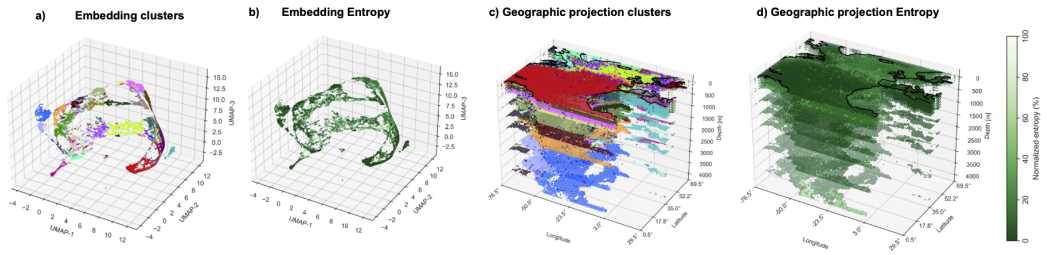


Figure 6: **Ocean biogeochemistry.** a) Embedding-space clusters, and b) embedding-space entropy from a representative ensemble member. c) The same clusters projected back to physical coordinates (depth, latitude, longitude), and d) Domain-space entropy field H_i. Regions ambiguous in latent geometry (high entropy in b) do not necessarily project to ambiguous regime assignments in domain space d), and vice versa.

the contrast across the Mediterranean outflow. Localized high entropy in regions where physical reasoning predicts smooth gradients is therefore a signal that the data is inadequate to constrain the ensemble, and can provide specific guidance for prioritizing future data collection.

6 Discussion and Limitations

NEMI reframes stochasticity in representation learning as a validation resource, with variability that single-run methods suppress becomes the empirical basis for assessing which discovered structures are robust to representation choice. The distinction from soft and spectral methods is where uncertainty is evaluated, not whether an ensemble is used. Soft clustering on a fixed embedding answers "given this representation, how confident are we in the assignment?" NEMI answers "given that the representation is itself uncertain, where in domain space is the regime structure ambiguous?" Methods on a single fixed representation provide embedding-space validation only, and their domain-space output is a deterministic projection of within-embedding uncertainty. The two questions above are complementary, and combining them by applying soft clustering within each ensemble member is left to future work. Spectral and diffusion-based methods could substitute for UMAP without modifying the validation principle.

A practical implication is that scientific conclusions should be conditioned on geographic uncertainty. Low-entropy regions support regime-based interpretation, and high-entropy regions remain scientifically meaningful, indicating overlap or ambiguity where a crisp assignment would mislead downstream inference. Low-entropy structure within expected boundary regions, in particular, generates testable hypotheses that single-realization clustering does not.

Limitations. Computational cost scales with ensemble size K, and effectiveness depends on sufficient sampling of the latent structure. Cluster correspondence under split/merge events is inherently ambiguous, and the spatial structure of H_i is preserved across alignment strategies, but per-cluster entropy values shift, and aggregate comparisons should be reported alongside the alignment strategy used. Lastly, aggregate entropy statistics treat grid cells as independent observations, not accounting for spatial autocorrelation, which is a known feature of geophysical fields and a topic of future work.

7 Conclusion

Stochasticity in nonlinear manifold learning, often treated as a problem, can be repurposed as a validation signal when projected into the domain space where scientific meaning is defined. We have formalized this principle as representation stability evaluated in domain space, demonstrated that the spatial structure of domain-space entropy is robust to alignment strategy across four Earth science applications, and shown that the geometry of where uncertainty concentrates is itself scientifically informative, identifying candidate refinements to established regime boundaries that single-realization clustering would obscure.

AI USE STATEMENT

In this work, we used generative AI tools to edit existing text (including shortening to meet page limits), find related work, to create BibTeX entries, and to provide critical feedback on a draft, and to produce this AI use statement. We have not used generative AI tools to develop the method, run analyses, or generate results. All AI-assisted text and references were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

AUTHOR CONTRIBUTIONS

Authors contributed in various levels, and will be detailed at a later stage in the review process to retain anonymity.

ACKNOWLEDGMENTS

Acknowledgments will be added at a later stage in the review process to retain anonymity.

REFERENCES

Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesús M. Pérez, and Iñigo Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, 2013. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2012.07.021. URL https://www.sciencedirect.com/science/article/pii/S003132031200338X.

Asa Ben-Hur, André Elisseeff, and Isabelle Guyon. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*, volume 7, pp. 6–17, 2002.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Stephen M Griffies, Alistair Adcroft, Rebecca Lynn Beadling, Mitchell Bushuk, Chiung-Yin Chang, Henri Francois Drake, Raphael Dussin, Robert W. Hallberg, William Hurlin, Hemant Khatri, John P Krasting, Matthew Lobo, Graeme MacGilchrist, Brandon G Reichl, Aakash Sane, Olga V. Sergienko, Maike Sonnewald, Jacob M. Steinberg, Jan-Erik Tesdal, Matthew D Thomas, Katherine Elise Turner, Marshall L Ward, Michael Winton, Niki Zadeh, Laure Zanna, Rong Zhang, Wenda Zhang, and Ming Zhao. The gfdl-cm4x climate model hierarchy, part i: model description and thermal properties. November 2024a. doi: 10.22541/essoar.173282145.53065190/v1. URL http://dx.doi.org/10.22541/essoar.173282145.53065190/v1.

Stephen M Griffies, Alistair Adcroft, Rebecca Lynn Beadling, Mitchell Bushuk, Chiung-Yin Chang, Henri Francois Drake, Raphael Dussin, Robert W. Hallberg, William Hurlin, Hemant Khatri, John P Krasting, Matthew Lobo, Graeme MacGilchrist, Brandon G Reichl, Aakash Sane, Olga V. Sergienko, Maike Sonnewald, Jacob M. Steinberg, Jan-Erik Tesdal, Matthew D Thomas, Katherine Elise Turner, Marshall L Ward, Michael Winton, Niki Zadeh, Laure Zanna, Rong Zhang, Wenda Zhang, and Ming Zhao. The gfdl-cm4x climate model hierarchy, part ii: case studies. November 2024b. doi: 10.22541/essoar.173282138.89550323/v1. URL http://dx.doi.org/10.22541/essoar.173282138.89550323/v1.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean Noël Thépaut. Era5 hourly data on pressure levels from 1979 to present. copernicus climate change service (c3s) climate data store (cds)., 2018. ISSN 1477870X.

Lawrence J. Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985. URL https://api.semanticscholar.org/CorpusID:189915041.

Yvonne Jenniges, Maike Sonnewald, Sebastian Maneth, Are Olsen, and Boris P. Koch. Unveiling 3d ocean biogeochemical provinces in the north atlantic: A systematic comparison and validation of clustering methods, 2025. URL https://arxiv.org/abs/2504.18181.

Hemant Khatri, Stephen M Griffies, Benjamin A Storer, Michele Buzzicotti, Hussein Aluie, Maike Sonnewald, Raphael Dussin, and Andrew Shao. A scale-dependent analysis of the barotropic vorticity budget in a global ocean simulation. *Journal of Advances in Modeling Earth Systems*, 16 (6):e2023MS003813, 2024.

Alexander Korablev, Are Olsen, Geophysical Institute, University of Bergen, and Bjerknes Centre for Climate Research. Comfort dataset, 2022. URL https://archive.sigma2.no//dataset/C31E1EFC-CE20-4357-BDA5-4DF32A3B792F.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52, 1955. URL https://api.semanticscholar.org/CorpusID:9426884.

Tilman Lange, Volker Roth, Mikio L. Braun, and Joachim M. Buhmann. Stability-based validation of clustering solutions. *Neural Computation*, 16(6):1299–1323, 2004. doi: 10.1162/089976604773717621.

Christopher D Manning. *Introduction to information retrieval*. Syngress Publishing,, 2008.

Makayla McDevitt, Maike Sonnewald, and Stephanie Dutkiewicz. From objective discovery to prediction of global ocean eco-provinces: A pathway for trustworthy learning, 2026. URL https://arxiv.org/abs/2609.13206.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3:861, 09 2018. doi: 10.21105/joss.00861.

Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1–2):91–118, 2003. doi: 10.1023/A:1023949509487.

Georg Stefan Schlake and Christian Beecks. Validating arbitrary shaped clusters - a survey. In *2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–12, 2024. doi: 10.1109/DSAA61799.2024.10722773.

Maike Sonnewald. A hierarchical ensemble manifold methodology for new knowledge on spatial data: An application to ocean physics. *Artificial Intelligence for the Earth Systems*, April 2026. doi: 10.22541/essoar.168056792.25480169/v1. URL http://dx.doi.org/10.22541/essoar.168056792.25480169/v1.

Maike Sonnewald, Carl Wunsch, and Patrick Heimbach. Unsupervised learning reveals geography of global ocean dynamical regions. *Earth and Space Science*, 6(5):784–794, 2019.

Maike Sonnewald, Stephanie Dutkiewicz, Christopher Hill, and Gael Forget. Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces. *Science advances*, 6 (22):eaay4740, 2020.

Maike Sonnewald, Redouane Lguensat, Dani Jones, Peter Düben, Julien Brajard, and V Balaji. Bridging observations, theory and numerical simulation of the ocean using machine learning. *Environmental Research Letters*, 16, 07 2021. doi: 10.1088/1748-9326/ac0eb0.

Maike Sonnewald, Krissy Anne Reeve, and Redouane Lguensat. A southern ocean supergyre as a unifying dynamical framework identified by physics-informed machine learning. *Communications Earth & Environment*, 4(1):153, 2023.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

Ulrike von Luxburg. Clustering stability: An overview. *Foundations and Trends in Machine Learning*, 2(3):235–274, 2010. doi: 10.1561/2200000008.

A PER-CLUSTER ALIGNMENT-SENSITIVITY RESULTS

This appendix provides per-cluster breakdowns of the aggregate alignment-sensitivity results reported in Table 1. For each domain, we report the mean and standard deviation of the entropy field H_i within each cluster, under both Hungarian and Greedy alignment strategies. The Δ Mean column reports the per-cluster difference (Greedy $-$ Hungarian), with negative values indicating lower entropy under Greedy.

A.1 OCEAN ECOLOGY

Table 2 reports per-cluster entropy for the ocean ecology application. Two clusters (8 and 9) containing trivially few points produced degenerate statistics under one or both strategies and are excluded. Across the remaining eight clusters, Greedy yields substantially lower entropy in clusters subject to split/merge across ensemble realizations (clusters 0, 4, 5), while Hungarian yields lower entropy in clusters where one-to-one correspondence is well-defined (clusters 3, 7). Differences in clusters 2 and 6 are within one percentage point.

	Hungarian		Greedy		
Cluster	Mean	Std	Mean	Std	Δ Mean
0	39.13	13.00	20.65	15.57	-18.49
1	27.61	10.31	22.69	7.50	-4.92
2	25.79	5.10	26.98	7.46	$+1.19$
3	13.21	12.51	25.56	2.74	$+12.35$
4	32.32	13.23	22.44	10.67	-9.88
5	31.53	7.36	21.04	8.94	-10.49
6	34.51	6.94	35.07	15.96	$+0.56$
7	26.43	5.05	30.00	7.09	$+3.57$
Mean	28.81	9.19	25.55	9.99	-3.26

Table 2: Per-cluster mean and standard deviation of entropy (%) for the ocean ecology application under Hungarian and Greedy alignment.

A.2 OCEAN DYNAMICS

Tables 3 and 4 report per-cluster entropy for the ocean dynamics application at 6 and 15 complexity levels. At 6, Hungarian and Greedy each yield lower entropy in three of six clusters, with the largest per-cluster differences (cluster 4 favoring Hungarian by 21.21 percentage points; cluster 1 favoring Greedy by 11.12) indicating regimes that split or merge across realizations. At 15, per-cluster differences shrink in magnitude and Hungarian wins more often (9 of 15 clusters), consistent with finer partitioning producing more clusters with stable one-to-one correspondence across realizations.

Table 3: Per-cluster mean and standard deviation of entropy (%) for the ocean dynamics application at 6, under Hungarian and Greedy alignment. The aggregate weighted mean (bottom row) matches Table 1.

	Hungarian		Greedy		
Cluster	Mean	Std	Mean	Std	Δ Mean
1	42.15	16.79	31.03	21.90	-11.12
2	36.08	19.76	29.69	23.61	-6.39
3	24.82	24.80	27.65	27.20	+2.83
4	32.80	19.71	54.01	13.56	+21.21
5	34.39	16.91	35.27	12.21	+0.88
6	41.91	16.11	40.69	14.47	-1.22
Mean	32.07	19.37	35.88	22.02	+3.81

Table 4: Per-cluster mean and standard deviation of entropy (%) for the ocean dynamics application at 15, under Hungarian and Greedy alignment. The aggregate weighted mean (bottom row) matches Table 1.

	Hungarian		Greedy		
Cluster	Mean	Std	Mean	Std	Δ Mean
1	29.70	13.57	31.03	16.87	+1.33
2	29.10	15.20	30.28	15.89	+1.18
3	37.88	12.29	41.70	11.12	+3.82
4	37.26	12.29	36.56	13.77	-0.71
5	23.13	23.39	18.91	19.63	-4.22
6	23.81	12.61	23.55	12.34	-0.26
7	34.84	16.01	24.67	14.75	-10.17
8	15.19	15.42	17.90	18.40	+2.71
9	18.91	16.40	18.76	13.28	-0.15
10	32.99	12.81	35.45	10.68	+2.46
11	40.91	15.30	38.91	12.62	-2.00
12	21.55	16.06	27.09	13.44	+5.54
13	16.59	18.17	27.69	11.22	+11.10
14	47.13	11.33	57.13	6.77	+10.00
15	44.24	6.82	51.94	4.46	+7.70
Mean	28.24	17.53	29.25	16.98	+1.01

A.3 Ocean biochemistry

Table 5: Per-cluster mean and standard deviation of entropy (%) for the ocean biogeochemistry application under Hungarian and Greedy alignment, with Δ Mean = Greedy − Hungarian. Hungarian alignment yields lower entropy in 246 of 310 clusters; the aggregate weighted means (8.53% Hungarian, 19.75% Greedy) are reported in Table 1.

		Hungarian		Greedy		
Cluster	N	Mean	Std	Mean	Std	Δ Mean
0	4978	3.63	8.69	47.97	1.49	+44.34
1	2650	9.69	13.75	9.94	14.47	+0.25
2	2640	8.09	11.21	23.13	8.88	+15.04
3	2299	16.34	6.41	28.86	10.89	+12.51
4	2309	3.98	7.05	8.30	6.76	+4.32
5	2806	12.13	8.76	32.61	5.85	+20.49
6	2100	0.73	4.23	0.76	4.37	+0.03
7	2083	3.60	5.37	20.81	4.82	+17.21
8	1927	13.01	3.81	29.29	3.04	+16.28
9	1891	9.11	10.50	14.18	12.89	+5.07
10	1769	21.82	9.35	40.31	4.31	+18.49
11	1547	4.25	5.62	5.89	5.57	+1.64
12	1373	3.79	8.53	3.80	8.41	+0.01
13	1206	4.89	10.68	4.92	10.71	+0.02
14	1155	15.82	5.07	35.91	3.64	+20.10
15	974	4.23	7.80	5.39	7.50	+1.16
16	987	16.38	15.11	17.01	15.76	+0.63
17	985	2.88	8.02	3.00	8.36	+0.12
18	871	2.64	9.29	2.72	9.62	+0.08
19	943	17.77	11.35	16.62	13.17	-1.15
20	584	13.56	8.44	10.57	9.60	-3.00
21	537	10.32	4.92	11.47	5.04	+1.15

(continued on next page)

(Table 5 continued)

		Hungarian		Greedy		
Cluster	N	Mean	Std	Mean	Std	Δ Mean
22	446	7.88	2.72	29.29	1.89	+21.41
23	424	0.25	1.91	0.26	2.02	+0.01
24	419	3.91	5.27	20.89	7.60	+16.97
25	315	27.82	20.34	42.57	11.22	+14.74
26	295	15.54	8.41	15.55	7.70	+0.01
27	19	26.62	1.21	39.11	0.53	+12.49
28	47	26.48	3.71	32.84	2.83	+6.36
29	278	19.43	3.33	40.00	1.80	+20.57
30	254	0.37	1.61	0.51	2.51	+0.14
31	258	5.02	2.71	5.95	2.72	+0.93
33	219	19.18	3.13	34.85	4.27	+15.68
34	235	13.48	4.39	16.68	22.07	+3.20
35	173	0.38	0.57	0.38	0.57	-0.00
36	172	1.41	2.17	2.73	3.25	+1.32
37	155	5.40	14.30	5.69	15.06	+0.30
38	151	4.82	6.96	4.86	7.07	+0.04
39	148	17.44	8.52	21.10	10.73	+3.66
40	139	5.83	11.31	7.20	10.85	+1.37
41	120	1.71	3.46	1.69	3.40	-0.01
42	108	26.59	7.76	36.64	1.76	+10.05
43	107	21.63	20.19	19.91	25.80	-1.72
44	107	4.02	3.43	5.12	3.61	+1.10
45	103	6.54	14.56	7.00	13.61	+0.45
46	109	7.19	6.92	10.20	13.02	+3.01
47	102	12.46	17.73	12.45	18.59	-0.01
48	104	22.99	6.11	24.00	6.85	+1.01
49	150	4.23	5.63	5.97	5.62	+1.74
50	93	17.31	9.70	37.53	5.25	+20.22
51	95	1.34	1.86	1.36	1.95	+0.02
52	93	10.64	6.37	12.09	5.48	+1.45
53	93	0.28	1.23	0.30	1.34	+0.02
54	90	21.91	20.29	24.28	18.72	+2.37
55	90	12.34	4.92	14.71	4.34	+2.37
56	90	4.68	3.38	12.79	4.18	+8.11
57	85	0.81	2.33	2.38	2.09	+1.57
58	82	15.65	8.08	17.75	8.02	+2.10
59	81	10.19	1.36	11.95	1.38	+1.76
60	78	9.46	11.74	11.73	11.12	+2.27
61	13	20.34	1.44	37.34	1.22	+17.00
62	70	2.33	4.88	11.42	3.90	+9.09
63	66	14.99	16.78	19.68	16.11	+4.69
64	80	3.26	4.74	8.57	4.63	+5.31
65	68	4.97	9.25	11.75	9.05	+6.78
66	64	5.18	3.32	9.57	3.40	+4.39
67	63	7.15	2.29	3.83	3.07	-3.32
68	63	2.07	2.80	5.56	2.80	+3.49
69	60	0.18	0.91	2.05	0.90	+1.87
70	83	7.44	5.28	13.54	4.57	+6.10
71	57	6.07	2.92	9.62	2.93	+3.55
72	55	2.52	5.29	3.50	5.44	+0.98
74	56	0.05	0.38	0.05	0.38	+0.00
75	57	7.90	0.63	7.91	0.70	+0.02
76	54	5.09	4.97	5.08	4.93	-0.01

(continued on next page)

(Table 5 continued)

		Hungarian		Greedy		
Cluster	N	Mean	Std	Mean	Std	Δ Mean
77	52	0.25	1.26	0.28	1.41	+0.03
78	52	3.58	0.88	0.26	0.89	-3.32
79	53	0.89	2.51	0.95	2.69	+0.06
80	49	18.48	1.66	34.54	1.41	+16.05
81	49	0.00	0.00	0.00	0.00	+0.00
82	49	10.55	1.38	12.11	1.48	+1.56
83	47	0.67	1.40	4.03	1.47	+3.36
84	48	1.32	1.62	2.30	1.81	+0.98
86	47	4.14	3.58	4.20	3.80	+0.06
87	46	10.23	0.00	11.84	0.00	+1.61
88	40	14.50	0.82	15.63	0.82	+1.13
89	46	14.41	2.13	31.25	2.05	+16.84
90	44	5.49	2.44	5.46	2.67	-0.03
91	43	1.86	0.40	2.79	0.40	+0.93
92	43	10.56	17.93	11.16	17.25	+0.60
93	43	0.00	0.00	0.00	0.00	+0.00
94	41	7.88	4.01	7.89	4.02	+0.01
95	62	9.29	3.00	15.05	2.52	+5.76
97	40	0.11	0.67	4.76	0.59	+4.66
99	41	4.84	8.57	5.71	9.85	+0.88
100	37	5.94	1.72	7.99	1.63	+2.04
101	36	0.94	0.00	0.00	0.00	-0.94
102	36	1.61	1.45	3.25	1.44	+1.63
103	39	9.11	3.25	11.24	4.55	+2.13
104	35	0.07	0.32	0.07	0.32	+0.00
105	31	21.47	10.69	34.67	7.39	+13.20
106	34	0.03	0.16	0.03	0.16	+0.00
107	33	14.94	1.11	14.96	1.22	+0.02
108	33	10.89	2.24	16.50	2.42	+5.61
109	31	4.97	3.43	7.46	3.21	+2.49
110	32	0.84	2.84	0.92	3.16	+0.08
111	32	0.65	1.91	0.70	2.12	+0.05
112	30	6.72	0.47	23.72	0.38	+17.00
113	27	11.23	0.79	11.23	0.79	+0.00
114	28	7.52	4.53	10.30	4.26	+2.77
116	28	4.01	4.41	4.08	4.53	+0.08
117	26	6.67	2.85	6.60	2.85	-0.07
119	27	11.74	0.61	12.58	0.69	+0.84
120	26	1.95	1.55	1.97	1.67	+0.02
121	26	11.60	1.65	11.54	1.65	-0.06
122	25	0.69	2.25	0.76	2.59	+0.07
123	24	25.96	1.19	37.97	2.27	+12.00
125	29	4.66	6.04	6.66	10.53	+2.00
127	24	0.04	0.19	0.04	0.19	+0.00
128	23	9.61	6.22	9.29	6.49	-0.32
130	22	2.57	2.17	2.80	2.17	+0.23
131	22	10.56	2.48	10.59	2.59	+0.03
132	25	10.08	16.76	9.85	16.12	-0.22
134	19	11.64	0.37	11.64	0.37	+0.00
135	21	4.91	4.09	5.22	2.93	+0.32
136	19	1.77	3.55	1.81	3.62	+0.04
137	17	1.54	4.11	1.86	5.10	+0.31
138	20	4.08	3.54	4.13	3.75	+0.05

(continued on next page)

(Table 5 continued)

		Hungarian		Greedy		
Cluster	N	Mean	Std	Mean	Std	Δ Mean
139	18	16.21	20.36	16.74	20.86	+0.53
140	19	2.40	2.89	1.75	3.54	-0.65
141	20	7.94	1.67	7.94	1.67	+0.00
142	18	9.57	1.49	9.57	1.49	+0.00
143	17	4.06	2.80	4.95	2.74	+0.89
144	5	7.89	0.60	62.79	0.24	+54.90
145	17	4.14	3.91	4.21	3.94	+0.08
146	17	4.11	4.05	5.37	4.94	+1.26
147	16	0.10	0.41	0.10	0.41	+0.00
148	16	22.85	6.14	34.59	6.18	+11.74
149	18	4.20	3.99	3.86	4.49	-0.34
150	15	17.22	1.25	17.22	1.25	+0.00
151	14	12.09	0.41	14.07	0.40	+1.97
152	15	12.80	1.76	12.68	2.08	-0.12
153	14	9.68	6.32	15.30	5.88	+5.62
154	15	9.52	1.77	15.85	1.60	+6.33
155	15	0.49	1.90	0.49	1.90	+0.00
156	15	1.35	2.94	1.45	3.32	+0.10
157	4	7.38	0.00	15.14	0.00	+7.75
158	13	12.02	1.82	12.14	2.09	+0.11
159	13	6.91	4.09	6.78	3.92	-0.13
160	13	7.87	0.47	7.87	0.47	+0.00
161	13	0.78	1.44	0.78	1.44	+0.00
162	13	0.00	0.00	0.00	0.00	+0.00
163	13	9.03	0.00	9.03	0.00	+0.00
164	13	4.89	2.74	5.44	3.00	+0.55
165	13	12.60	2.82	13.46	2.81	+0.86
166	13	1.41	1.70	12.22	1.43	+10.82
167	14	2.98	4.79	3.14	5.16	+0.16
169	14	3.59	2.47	8.25	2.49	+4.66
170	12	6.30	3.50	5.46	3.59	-0.84
171	12	36.36	9.41	36.02	11.22	-0.34
172	12	0.14	0.47	0.14	0.47	+0.00
173	12	11.60	0.00	16.26	0.00	+4.66
174	12	0.31	0.46	1.95	0.46	+1.64
175	12	18.41	7.25	33.85	4.87	+15.45
176	13	11.33	3.38	11.46	3.66	+0.13
177	32	19.32	3.42	19.64	3.60	+0.32
178	12	1.64	0.00	1.64	0.00	+0.00
179	11	12.96	1.71	27.96	1.13	+15.00
180	11	2.71	3.24	2.82	3.62	+0.12
181	11	3.90	3.62	3.97	3.85	+0.07
182	6	13.01	3.40	17.05	1.86	+4.04
183	12	11.80	2.92	11.99	3.33	+0.19
184	11	1.09	0.49	2.02	0.49	+0.94
186	13	4.53	5.59	5.69	8.58	+1.17
188	9	0.00	0.00	0.00	0.00	+0.00
189	10	8.20	0.71	8.23	0.80	+0.03
190	9	0.10	0.31	0.10	0.31	+0.00
191	10	14.08	3.54	17.39	2.87	+3.32
192	10	11.19	0.52	13.65	0.41	+2.45
193	9	10.23	0.70	14.20	0.72	+3.97
194	9	9.04	1.31	39.60	0.63	+30.56

(continued on next page)

(Table 5 continued)

		Hungarian		Greedy		
Cluster	N	Mean	Std	Mean	Std	Δ Mean
195	9	1.77	3.64	1.91	4.03	+0.14
196	11	9.87	4.13	13.11	3.96	+3.24
197	9	1.04	0.31	8.76	0.29	+7.72
199	8	2.14	2.76	2.21	2.83	+0.06
200	9	3.09	2.72	3.01	0.91	-0.07
201	52	44.10	11.66	41.41	10.61	-2.70
202	9	1.02	1.68	1.02	1.68	+0.00
203	9	1.15	0.41	1.15	0.41	+0.00
204	10	16.95	1.07	9.72	2.53	-7.23
205	9	1.82	0.54	2.44	0.54	+0.61
206	9	0.77	1.64	0.83	1.79	+0.06
207	9	0.00	0.00	0.00	0.00	+0.00
209	8	4.35	0.99	4.35	0.99	+0.00
210	7	14.66	0.44	28.97	0.00	+14.31
212	8	0.12	0.33	0.12	0.33	+0.00
213	7	7.71	2.47	7.71	2.47	+0.00
214	8	0.12	0.33	0.12	0.33	+0.00
215	8	3.86	2.43	4.56	2.42	+0.70
216	8	0.00	0.00	0.00	0.00	+0.00
217	8	0.94	0.00	1.64	0.00	+0.70
218	8	0.94	0.00	0.94	0.00	+0.00
219	8	13.41	0.00	13.41	0.00	+0.00
220	8	7.37	0.61	7.37	0.61	+0.00
222	11	23.39	17.30	27.41	21.56	+4.02
223	7	11.42	1.81	11.55	2.15	+0.13
224	6	17.75	1.54	31.06	0.90	+13.31
225	7	1.19	2.37	1.19	2.37	+0.00
226	7	13.48	3.16	20.24	1.84	+6.76
227	9	19.88	2.75	20.32	3.69	+0.44
228	7	3.72	0.57	23.27	1.67	+19.54
229	7	4.05	1.56	5.05	1.73	+1.00
230	13	9.64	5.40	15.48	11.24	+5.85
231	8	4.87	0.57	35.79	0.39	+30.92
232	7	7.37	1.84	7.37	1.84	+0.00
233	7	0.23	0.62	0.23	0.62	+0.00
234	6	7.90	3.32	12.91	3.12	+5.01
235	11	18.29	2.54	18.05	6.96	-0.25
237	7	6.94	3.56	7.66	3.78	+0.72
238	7	0.00	0.00	0.00	0.00	+0.00
239	7	12.18	0.00	16.19	0.00	+4.01
240	6	3.05	3.19	4.79	3.11	+1.74
241	7	9.42	1.60	10.06	1.60	+0.64
242	6	0.31	0.76	0.31	0.76	+0.00
243	6	2.98	4.58	12.91	0.97	+9.93
244	5	0.64	0.99	8.26	0.47	+7.62
245	6	6.86	4.20	12.76	7.74	+5.91
246	6	9.28	0.00	18.72	0.00	+9.44
247	6	2.03	3.97	4.06	2.88	+2.03
248	7	3.46	3.80	3.57	4.10	+0.11
249	7	3.09	5.28	4.80	4.99	+1.70
251	3	17.52	0.32	26.91	0.55	+9.39
252	6	2.38	3.63	4.27	3.03	+1.89
253	6	12.43	0.00	12.55	0.00	+0.12

(continued on next page)

(Table 5 continued)

		Hungarian		Greedy		
Cluster	N	Mean	Std	Mean	Std	Δ Mean
254	7	20.69	1.71	25.67	1.07	+4.98
255	5	14.83	0.00	15.67	0.00	+0.83
257	6	1.47	0.92	1.99	0.29	+0.52
258	6	0.16	0.38	0.16	0.38	+0.00
259	6	1.09	0.38	2.03	0.38	+0.94
260	5	15.49	1.46	21.92	0.58	+6.43
261	6	1.37	0.70	2.15	0.43	+0.78
262	7	6.76	3.48	10.91	4.07	+4.15
263	6	12.92	1.96	27.98	3.85	+15.06
264	6	6.09	5.40	9.30	4.19	+3.21
265	6	0.00	0.00	0.00	0.00	+0.00
266	6	10.00	4.38	18.55	2.16	+8.56
268	6	8.66	0.30	8.77	0.57	+0.11
269	6	13.66	2.86	18.76	1.41	+5.11
270	6	1.40	3.42	4.62	2.16	+3.23
271	8	21.42	3.28	41.44	1.08	+20.03
272	5	1.07	2.40	17.13	0.79	+16.06
273	4	1.64	0.00	26.84	0.00	+25.20
274	5	16.11	1.20	26.52	0.26	+10.41
275	5	0.00	0.00	20.98	0.00	+20.98
276	5	2.01	0.31	15.16	0.04	+13.15
277	5	6.09	1.42	24.98	0.07	+18.90
278	7	22.23	13.46	28.68	7.21	+6.44
279	5	11.04	0.49	23.25	0.47	+12.21
280	8	10.53	5.03	34.42	4.89	+23.89
281	5	13.09	1.32	28.79	0.77	+15.70
282	5	7.33	2.16	25.28	1.13	+17.95
283	6	2.22	5.44	9.28	2.80	+7.06
285	5	2.19	2.34	33.46	0.84	+31.27
286	5	11.92	3.37	37.11	3.38	+25.19
287	1	22.24	nan	29.38	nan	+7.14
288	4	12.85	0.69	32.83	0.82	+19.97
289	1	20.28	nan	29.79	nan	+9.50
290	4	13.67	0.75	43.73	0.52	+30.05
292	5	8.98	0.40	40.71	0.34	+31.73
293	7	10.62	11.80	36.64	5.99	+26.02
294	6	15.09	0.94	24.81	1.75	+9.71
295	6	7.15	3.08	27.25	1.75	+20.10
297	4	14.73	0.62	46.58	0.42	+31.85
299	9	2.82	1.80	57.82	6.35	+55.00
301	6	15.15	2.57	37.26	0.74	+22.11
302	5	0.19	0.42	54.66	0.00	+54.48
303	4	10.15	0.21	50.86	0.24	+40.71
304	8	18.09	4.38	28.82	1.96	+10.73
305	5	9.76	1.28	39.06	0.36	+29.30
306	5	4.32	1.73	54.60	0.88	+50.28
307	7	9.04	3.96	23.00	1.90	+13.97
311	4	3.33	0.00	35.87	0.00	+32.55
314	4	9.69	0.26	56.03	0.21	+46.33
315	6	1.56	1.84	41.67	0.30	+40.11
316	4	7.63	0.00	53.53	0.00	+45.90
322	4	1.64	0.00	50.06	0.00	+48.42
325	3	18.33	0.05	49.18	0.04	+30.85

(continued on next page)

(Table 5 continued)

		Hungarian		Greedy		
Cluster	N	Mean	Std	Mean	Std	Δ Mean
327	7	2.87	2.83	49.75	2.44	+46.89
329	4	6.85	1.48	54.59	1.03	+47.73
330	4	14.36	2.65	58.16	0.19	+43.80
331	4	10.90	0.09	31.61	0.37	+20.71
333	4	0.95	0.00	45.41	0.00	+44.46
335	4	5.97	0.33	37.90	0.27	+31.93
337	4	8.00	1.46	56.89	0.38	+48.88
340	4	3.94	0.22	41.70	0.11	+37.76
342	4	7.12	0.00	27.85	0.00	+20.73
347	4	0.00	0.00	51.08	0.00	+51.08
351	3	9.28	0.65	49.81	0.56	+40.53
354	4	9.28	0.45	50.47	0.35	+41.20
356	4	13.48	0.00	40.73	0.00	+27.25
359	3	20.73	2.73	52.05	1.43	+31.32
360	4	13.30	1.25	62.68	0.40	+49.38
361	4	1.65	0.00	52.78	0.00	+51.12
366	4	0.00	0.00	54.93	0.00	+54.93
369	4	8.15	1.14	41.69	0.30	+33.53
375	4	7.05	0.00	39.37	0.00	+32.32
377	3	7.66	0.00	58.58	0.00	+50.91
378	4	2.87	0.58	50.98	0.03	+48.12
382	3	12.00	0.29	64.03	0.31	+52.03
384	3	15.38	0.00	59.21	0.00	+43.83

B Ocean plankton ecology and atmospheric dynamics data

Details here can be found in manuscripts under review and will be made available for a camera-ready version of the present manuscript.

C Ocean Physical Diagnostic Terms (Barotropic Vorticity)

The main body of the paper is domain-agnostic, as the validation principle and entropy-based overlap diagnostic apply broadly. This appendix documents the specific physically interpretable diagnostic terms used in the geophysical case study of ocean physics and provides context for the "term balance" summaries shown in Fig. 4.

C.1 Diagnostic Feature Construction

The data used for the ocean physics example come from the Modular Ocean Model version 6 (MOM6) in the CM4X Climate Model hierarchy at 1/4 degree ocean resolution (Griffies et al., 2024a;b)). The terms were computed after Khatri et al. (2024).

C.2 Barotropic Vorticity Terms

For our example, we used a vorticity budget based on the depth-integrated Boussinesq-hydrostatic ocean primitive equations. From Khatri et al. (2024): "The complete vorticity budget of the depth-integrated flow can be written as:

$$\beta V = \frac{J(p_b, H)}{\rho_o} + \hat{z}\cdot\left(\frac{\nabla\wedge\tau_s}{\rho_o} - \frac{\nabla\wedge\tau_b}{\rho_o} + \nabla\wedge\mathcal{A} + \nabla\wedge\mathcal{B}\right) - f\frac{Q_m}{\rho_o} + f\,\partial_t\eta - \hat{z}\cdot(\nabla\wedge\mathcal{U}_t)\,, \quad (9)$$

where $\beta = \partial_y f$ is the meridional derivative of the Coriolis parameter, V is the vertically-integrated meridional velocity, $z = \eta$ is the ocean free surface height, $z = -H$ is ocean bottom, p_b is bottom

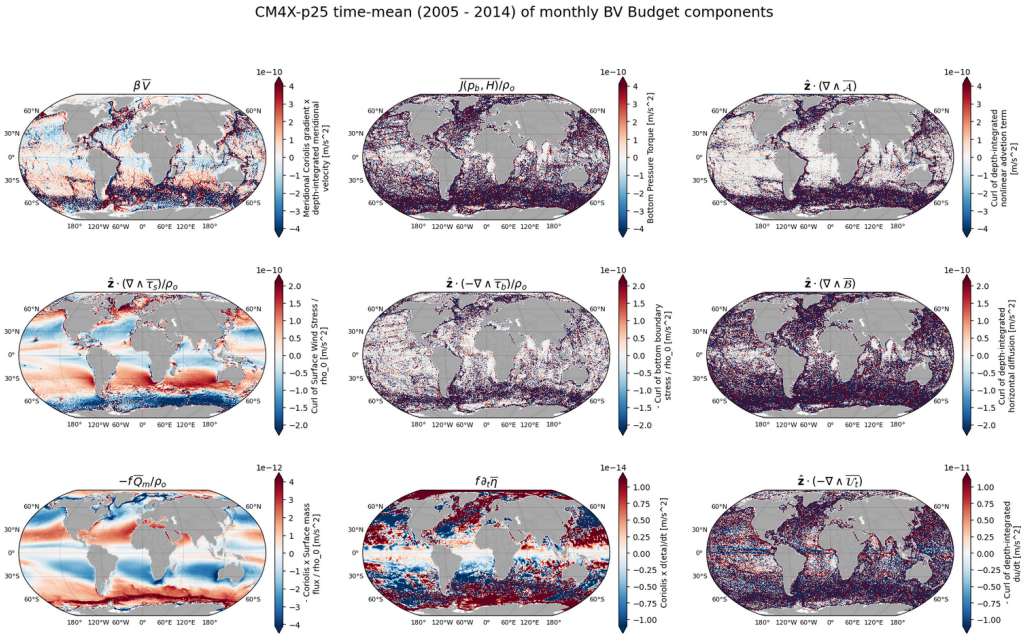


Figure 7: Barotropic vorticity diagnostic terms used in the case study. These terms provide physically interpretable inputs for unsupervised discovery and support post hoc interpretation of discovered regimes via cluster-wise term balances.

pressure, $\nabla = \hat{x}\,\partial_x + \hat{y}\,\partial_y$ is the horizontal gradient operator, and $\rho_o = 1035$ kg m^{-3} is the Boussinesq reference density. τ_s and τ_b are surface wind stress and bottom friction stress vectors, respectively. $\mathcal{A}$ and $\mathcal{B}$ represent the vertically integrated velocity advection and horizontal viscous friction terms. Here, Q_m is the downward mass flux on the ocean surface, and $\mathcal{U}_t$ is the vertically integrated velocity tendency term. By assuming a steady state, negligible horizontal friction, $Q_m = 0$, linearity, and a flat bottom ocean, equation (9) readily reduces to the Stommel model of wind-driven gyre given by the windstress curl balancing $\beta\, V$. The terms in Eq. 10 correspond to the bars in Fig. 4b and e.