https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

CMIP6 biogeochemical models under-represent the deep chlorophyll maximum and diverge in how they express it: a regime-stratified BGC-Argo evaluation in the North Pacific

Taketo Hashioka[1] and Maki N. Aita[1]

[1] Research Institute for Global Change, Japan Agency for Marine-Earth Science and Technology, 2-15, Natsushima-cho, Yokosuka, Kanagawa, 237-0061, Japan

Correspondence to: Taketo Hashioka (hashioka@jamstec.go.jp)

Abstract

Surface chlorophyll, the variable most directly linked to satellite ocean colour, is widely used to evaluate marine biogeochemical models, yet it cannot by itself reveal how chlorophyll is distributed through the water column. Using Biogeochemical-Argo (BGC-Argo) chlorophyll profiles over the North Pacific, we evaluate the vertical chlorophyll structure of nine Coupled Model Intercomparison Project Phase 6 (CMIP6) biogeochemical models, and we ask how faithfully the models reproduce that structure — including where modelled and observed surface chlorophyll already agree. The observations show that the deep chlorophyll maximum (DCM; median depth 71 m) is not a single feature but a family of vertical structures: grouped by the observed DCM depth, the observed maximum both deepens and becomes more sharply expressed, from shallow and weakly expressed to deep and strongly expressed. Relative to these observations, the models allocate a smaller fraction of chlorophyll to the subsurface and express a weaker vertical chlorophyll contrast (DCM prominence) — the subsurface-fraction bias is negative and the prominence ratio below unity in all nine models across the full comparison — and they place the chlorophyll maximum increasingly too shallow as the observed DCM deepens, with a median DCM-depth bias of −9, −22, and −37 m across the shallow, intermediate, and deep observed-DCM regimes. The models further separate into distinct types of DCM representation — flat, shallow, and deep — none of which reaches the observed combination of DCM depth and prominence, so they differ in kind and not only in degree. Because the models place their maxima at different depths, the ensemble median profile washes the maximum out and is a poor summary of its own members; ensemble-level agreement in DCM depth would therefore reflect cancellation among per-model errors of opposing sign rather than per-model fidelity. These vertical-allocation biases persist even where modelled and observed surface chlorophyll agree: within the surface-matched subset the median subsurface chlorophyll fraction bias is about −5 to −6 percentage points with an interquartile range entirely below

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

zero, and the DCM prominence ratio remains below unity (0.81). Surface chlorophyll agreement is therefore not sufficient to diagnose biases in vertical chlorophyll allocation, and surface-focused evaluation can leave systematic subsurface structural biases undetected. Diagnosing and improving the vertical distribution of chlorophyll in biogeochemical models benefits from depth-resolved constraints such as those provided by BGC-Argo.

1 Introduction

Marine biogeochemical models are central to projecting the ocean's role in carbon cycling and to understanding how marine ecosystems respond to climate change. Chlorophyll-a concentration is one of the most widely used variables for evaluating these models, because it is an observable proxy for phytoplankton and, at the surface, is directly linked to satellite ocean-colour measurements. Chlorophyll thereby occupies an unusual position among the observational constraints on ocean biogeochemistry: most of the others, nutrients among them, are available mainly as climatologies, whereas satellite ocean colour resolves chlorophyll in space and in time across whole basins. In consequence, model assessment and tuning have relied heavily on surface chlorophyll, and much of our confidence in simulated ocean biology rests on how well models reproduce it.

Surface chlorophyll, however, characterizes only the top of the water column. In large parts of the stratified ocean, chlorophyll is greatest not at the surface but at a subsurface maximum, the deep chlorophyll maximum (DCM), which forms below the depths sampled by ocean colour. A substantial fraction of water-column chlorophyll, and of the associated biological structure, therefore resides where satellites cannot see it. Two water columns with similar surface chlorophyll can differ markedly in the depth, intensity, and shape of their subsurface chlorophyll, so surface chlorophyll alone provides an incomplete constraint on the vertical distribution of chlorophyll.

The vertical structure of chlorophyll, and the DCM in particular, has been documented extensively from ship-based observations and, more recently, at unprecedented coverage by Biogeochemical-Argo (BGC-Argo) profiling floats. These autonomous observations show that DCMs are widespread and that their depth and intensity vary regionally with light, nutrients, and mixing (Cornec et al., 2021; Barbieux et al., 2019). The DCM is therefore not a single feature with a characteristic depth, but a family of vertical structures whose depth and expression change systematically from one setting to another. In the North Pacific specifically, a companion analysis of the same BGC-Argo observations resolves multiple distinct DCM structures, documenting this diversity directly (Hashioka and Aita, 2026a). BGC-Argo provides a depth-resolved observational constraint on

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

that family which surface-based data cannot, and which is well suited to asking not only whether biogeochemical models produce a DCM, but whether they reproduce the range of DCM structures that is actually observed.

Previous evaluations of marine biogeochemical models, particularly those participating in CMIP, have commonly used surface, upper-ocean, or vertically integrated diagnostics, including satellite-based surface chlorophyll, nutrients, oxygen, air–sea CO2 fluxes, and primary production (e.g., Séférian et al., 2020; Kwiatkowski et al., 2020). More recent studies have evaluated DCM-related and vertical chlorophyll diagnostics in specific regions, for example, across Coupled Model Intercomparison Project Phase 6 (CMIP6) models in the Southern Ocean (Cheng et al., 2025). On the observational side, recent syntheses have mapped observed subsurface chlorophyll structure across the global ocean, including the North Pacific, and have noted that model evaluation has largely relied on surface chlorophyll (Yasunaka et al., 2022). What remains unclear is how faithfully current models represent the observed vertical structure across the range of structures that is observed: whether they follow the observed deepening and sharpening of the DCM, whether they differ from one another in kind or only in degree, and whether agreement in surface chlorophyll is a useful indicator of subsurface fidelity. The last question matters because BGC-Argo observations show that surface chlorophyll alone does not uniquely determine subsurface or column chlorophyll structure (Cornec et al., 2021; Tanner et al., 2024).

Here we use BGC-Argo chlorophyll profiles over the North Pacific to evaluate the vertical chlorophyll structure of nine CMIP6 biogeochemical models. We first characterize the observed structure and show that it is a family organized by DCM depth rather than a single profile. We then evaluate the models within observed-DCM regimes, so that their fidelity is assessed both where the observed maximum is weakly and where it is strongly expressed. We next summarize each model by the depth and the prominence of its own chlorophyll maximum, which separates the ensemble into distinct types of DCM representation rather than along a single axis of skill. Finally, we isolate the subset of model–cell comparisons in which surface chlorophyll already agrees with the observations, as a direct test of whether surface agreement is sufficient to constrain vertical structure, and we assess the robustness of that test to the choice of BGC-Argo chlorophyll product. The analysis is diagnostic rather than mechanistic: we characterize and localize the biases and provide a compact, reproducible set of diagnostics, and we reserve mechanistic attribution for future work.

2 Data and Methods

2.1 BGC-Argo observations

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

We used BGC-Argo chlorophyll-a profiles over the North Pacific. Adjusted chlorophyll and pressure (CHLA_ADJUSTED, PRES_ADJUSTED) were quality-controlled using their associated quality-control flags, retaining flags 1 or 2. The chlorophyll product was then reprocessed to a community-consistent chlorophyll definition, harmonizing profiles that carried the earlier factor-of-two correction, as described in Section 2.2. The analysis domain spans 2.5–57.5° N and 122.5° E–92.5° W, so that the sampled "North Pacific" includes cells on the equatorward flank of the subtropical gyre as well as the subarctic. Profiles were aggregated onto a regular 5° × 5° grid, chosen coarse enough that cells are typically sampled by many profiles (a median of 38 profiles per cell over the 225 occupied cells, rising to 44 over the 201 cells that enter the comparison; Sections 2.5 and 4.6): within each occupied cell we formed cell-level profiles by depth layer, and the observed composite structure was summarized as the cell median with its inter-cell interquartile range. The full observed population comprises 225 occupied 5° cells and 12,583 profiles. Because chlorophyll from fluorescence is interpreted as adjusted pigment structure rather than absolute phytoplankton biomass, and because absolute chlorophyll is uncertain across both observations and models, our diagnostics are based on relative and structural quantities (Section 2.4) rather than on absolute chlorophyll magnitude.

2.2 Chlorophyll reprocessing

Observed chlorophyll profiles were obtained from BGC-Argo. Profiles identified as carrying the earlier global factor-of-two correction (Roesler et al., 2017, for WET Labs ECO fluorometers) were rescaled to a community-consistent value using the local fluorescence-to-chlorophyll physiological ratio from the SEANOE global look-up table (Sauzède et al., 2026; DOI: 10.17882/105732), Version 2.0, quality-controlled NetCDF. For each 5° observed cell used in the model–observation comparison, we sampled the look-up table over the corresponding cell footprint and used the cell-median fluorescence-to-chlorophyll ratio, applied as a single depth-independent scale factor, to convert between the reprocessed and the older global factor-two definitions. We refer to the resulting product as the reprocessed BGC-Argo chlorophyll.

The reprocessing adjusts the absolute magnitude of chlorophyll, and because the correction is region-dependent, it acts in both directions across the sampled domain. It does not, however, alter the vertical position of chlorophyll features. In our data the observed DCM depth is identical before and after reprocessing — a per-cell, depth-independent rescaling cannot move the depth of the profile maximum — consistent with the community documentation (Section 2.4). Because the diagnostics that carry our main result are relative and structural — the surface-normalized profile shape, the subsurface chlorophyll fraction, and the DCM prominence, together with the DCM depth — they are robust to a multiplicative change in absolute chlorophyll. We accordingly frame our

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

results against a reprocessed BGC-Argo chlorophyll reference and do not interpret absolute chlorophyll differences. To confirm that this choice does not drive our conclusions, we recompute the surface-matched subsurface offset under both the reprocessed and the earlier factor-two definitions (Section 3.7, Fig. S1).

2.3 CMIP6 models

We evaluated nine CMIP6 biogeochemical models (Table 1). For each model we used the ocean chlorophyll field (chl, table Omon) from the historical experiment; the realization (variant label) and grid label differ among models and are listed explicitly in Table 1. The CMIP6 fields were selected according to source availability and provenance consistency for the required chl/Omon historical stream; no realization was selected on the basis of model performance. Native-grid model chlorophyll fields were remapped onto the common 5° × 5° horizontal grid using conservative remapping, and chlorophyll was then interpolated onto a common vertical (depth) grid spanning 0–300 m. Diagnostics were computed per profile and per month on the common depth grid, and annual values were defined as the mean of valid monthly diagnostics rather than as diagnostics of an annual-mean profile.

One model (CESM2) archives ocean chlorophyll only to approximately 145 m in its CMIP6 output, so on the common vertical grid its chlorophyll is resolved through the 125 m layer (the 112.5–137.5 m layer) and is undefined at 150 m and below. To evaluate all nine models on an identical basis, the integrated subsurface-fraction diagnostic (Section 2.4) is therefore defined over the deepest depth layers that all nine models resolve; the depth of the chlorophyll maximum and the DCM prominence, which are extremum-based rather than integrated, are unaffected.

The comparison is framed as a structural, climatological assessment: the observed years and the model historical period do not coincide, and we interpret the diagnosed differences as climatological structural biases rather than as an interannual, time-matched comparison. We do not rank the models or reduce their performance to a single skill score; models are treated as an ensemble in which common and model-dependent components of the biases are distinguished.

Table 1. CMIP6 models evaluated.

Legend	CMIP6 source_id	Variant	Grid	Model-description reference
ACCESS	ACCESS-ESM1-5	r10i1p1f1	gn	Ziehn et al. (2020)
CESM2	CESM2	r4i1p1f1	gn	Danabasoglu et al. (2020)
CanOE	CanESM5-CanOE	r1i1p2f1	gn	Christian et al. (2022)

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

CMCC	CMCC-ESM2	r1i1p1f1	gn	Lovato et al. (2022)
GFDL	GFDL-ESM4	r1i1p1f1	gr	Dunne et al. (2020)
IPSL	IPSL-CM6A-LR	r2i1p1f1	gn	Boucher et al. (2020)
MIROC	MIROC-ES2L	r1i1p1f2	gr1	Hajima et al. (2020)
MRI	MRI-ESM2-0	r1i2p1f1	gn	Yukimoto et al. (2019)
UKESM	UKESM1-0-LL	r1i1p1f2	gn	Sellar et al. (2019)

2.4 Vertical chlorophyll diagnostics

From each observed and modelled profile we computed four diagnostics of vertical chlorophyll structure, all evaluated on the common depth grid. (i) Surface chlorophyll is the median chlorophyll over the near-surface layer (0–10 m). (ii) DCM depth is the depth of the profile-wise maximum chlorophyll within 0–300 m, taken directly with no smoothing, filling, or extrapolation; for profiles whose maximum occurs near the surface, this diagnostic returns a near-surface depth, so that the absence of a subsurface maximum is expressed as a shallow DCM depth rather than excluded.

(iii) The subsurface chlorophyll fraction is the layer-integrated chlorophyll over 25–137.5 m divided by the layer-integrated chlorophyll over 7.5–137.5 m, expressed as a percentage; both integrals are evaluated as sums of the cell-median layer chlorophyll weighted by layer thickness on the common depth grid. The lower bound of the numerator (25 m) excludes the near-surface layer, and the lower bound of the denominator (7.5 m) excludes the sparsely sampled 0–7.5 m surface layer. The upper bound (137.5 m) is the upper edge of the deepest depth layer resolved by all nine models; because CESM2 archives chlorophyll only to about 145 m (Section 2.3), integrating deeper would exclude it. The observed and modelled fractions are computed identically — same window, same layer boundaries, and same layer-integration operator — so the fraction is a within-column allocation metric whose window choice is a definition applied equally to observations and models. We verified that, for the eight models that resolve deeper than 137.5 m, extending the window to 175 m leaves the fraction essentially unchanged, and that in the model-minus-observation bias the window contribution cancels because it enters observations and models equally (Supplement).

(iv) DCM prominence is defined per profile as the maximum chlorophyll over 5–145 m divided by the surface chlorophyll (0–10 m median), computed identically for the model and the observation; a prominence window starting at 5 m is used because prominence is a vertical-contrast maximum metric rather than an integrated subsurface-allocation fraction. Because DCM prominence shares surface chlorophyll with the surface chlorophyll ratio used as an abscissa in Fig. 6 (a definitional coupling), we interpret only its offset within the surface-matched

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

band and not its slope (Section 3.6). Biases are defined as model minus observation for depth and fraction, and as the model-to-observation ratio for prominence. Only model–cell pairs for which all four diagnostics were valid were retained; profiles or cells with missing or invalid diagnostics were excluded rather than filled.

(v) For the profile-shape comparisons of Sections 3.3, 3.4, and 4.5 we additionally use surface-normalized profiles: each observed or modelled cell profile is divided by its own 0–10 m surface chlorophyll, so that what is compared is the vertical shape of the profile rather than its absolute magnitude. Within a given group of cells — all 201 comparison cells, or the cells of a single observed-DCM regime — the central tendency of these normalized profiles is summarized at each depth as the median across cells, and their spatial variability as the inter-cell interquartile range across cells. We use the median rather than the mean for these profile summaries, because the cells contributing to a group are not independent and the normalized profiles are strongly right-skewed at depth; the choice is stated explicitly because the two estimators differ appreciably for these distributions. The CMIP6 ensemble profile shown in Fig. 4 is the median across the nine per-model median profiles, and is therefore a summary of model medians rather than a pooled median over all model–cell pairs. Where the spatial variability of the profiles is compared between a model and the observations (Section 4.5), the inter-cell interquartile range of each field is reduced to a single number by integrating its width over 40–150 m — the depth interval that contains the observed maxima of all three regimes — and the model value is expressed as a ratio to the observed value obtained over the same cells and the same interval; the integration is performed on the normalized profiles themselves, on the linear scale on which they are compared, so that depths at which the observed spread is negligible contribute negligibly to the ratio. Because a normalized profile is a ratio to the surface value, its magnitude is sensitive to that value; quantitative statements about the strength of the chlorophyll maximum are therefore carried by the DCM prominence diagnostic (iv) rather than by the normalized profiles themselves.

2.5 Model–cell matching and stratification

Each model was matched to the observations at the level of occupied 5° cells, forming model–cell pairs, with each observed cell contributing one pair per model. Because pairs from the same cell share the same observation, we treat the pooled distributions descriptively — using medians and interquartile ranges, and reporting pair-based and cell-based summaries in parallel — rather than applying significance tests; where relevant we separate within-model from between-model variability. In a cell-based summary the cell value is the mean of that cell's model–cell pairs and the reported statistic is the median, with its interquartile range, over cells, so that every cell

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

contributes once irrespective of how many pairs it carries. The full set of valid comparisons comprises 1,809 model–cell pairs across 201 cells, balanced at 201 pairs per model.

The comparison cells were grouped into shallow, intermediate, and deep observed-DCM regimes by tertiles of their observed DCM depth. The tertile thresholds are the lower and upper terciles of the observed DCM depth over the 201 comparison cells, 54.1 and 96.6 m, and they divide the cells into three groups of 67, with median observed DCM depths of 42.2, 71.1, and 112.1 m; the spatial distribution of the three regimes is shown in Fig. A1. This stratification corresponds to the same 1,809 model–cell pairs across the 201 cells, 603 in each regime, and is the basis of the regime-stratified evaluation in Sections 3.3–3.5. The regimes are defined from the observations alone: only the observed DCM depth of each cell enters the grouping, and no model field is used, so the regimes are fixed targets against which every model is evaluated on identical cells. The statistical consequences of stratifying on an observed quantity that also enters one of the evaluated biases are examined in Section 4.6.

The surface-matched band is defined as the subset of pairs with |log2(model/observation surface chlorophyll ratio)| < 0.5 (263 pairs across 102 cells) and is used for the sufficiency test of Section 3.6.

The populations used here are nested, and each is labelled explicitly wherever it appears. The observed population comprises 225 occupied cells and 12,583 profiles; of these, 203 cells are sampled by at least five profiles, and 201 of those yield all four observed diagnostics and enter the model–observation comparison — the two excluded cells have zero observed surface chlorophyll, so the prominence ratio is undefined there, and every cell that passes has a counterpart in all nine models; the surface-matched band is a subset of the latter (102 of the 201 cells). Quantities reported for the full comparison (1,809 pairs across 201 cells) and for the surface-matched band (263 pairs across 102 cells) are therefore not interchangeable and are never pooled.

3 Results

3.1 The observed deep chlorophyll maximum is a family of vertical structures

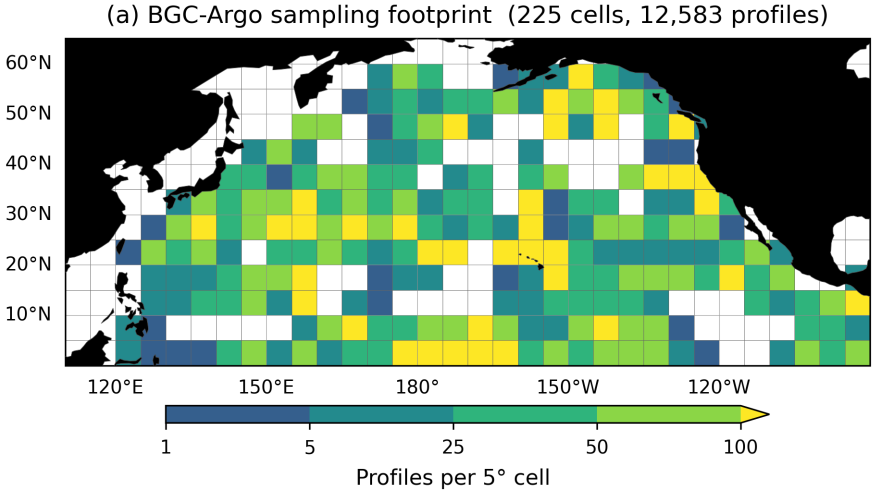
The BGC-Argo composite establishes the observed vertical chlorophyll structure that the models are evaluated against. Across the full observed population (225 five-degree cells, 12,583 profiles; BGC-Argo reprocessed chlorophyll), the North Pacific composite profile exhibits a pronounced deep chlorophyll maximum (DCM), with a median DCM depth of 71 m (Fig. 1). The presence and subsurface depth of this maximum are consistent with prior observation-based characterizations of the North Pacific deep chlorophyll maximum (Yasunaka et al.,

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

2022), supporting the observed composite as a basis for model evaluation. Chlorophyll is low in the surface layer, increases to a subsurface maximum near 60–70 m, and decreases below it, so that a substantial fraction of the water-column chlorophyll resides below the surface layer.

The composite alone, however, understates what a model has to reproduce. The observed DCM is not a single feature repeated across the basin but a family of vertical structures: its depth and intensity vary considerably among cells, as indicated by the wide inter-cell interquartile range around the composite (Fig. 1), and this variation is spatially organized rather than random (Fig. A1). Grouping the 201 comparison cells by tertiles of their observed DCM depth gives median observed DCM depths of 42.2, 71.1, and 112.1 m in the shallow, intermediate, and deep regimes (67 cells each; thresholds 54.1 and 96.6 m; Section 2.5). As shown in Section 3.4, the observed profile does not simply translate downwards across these regimes: where the observed DCM is deeper, it is also more sharply expressed relative to the surface. This diversity — and not the basin-scale composite by itself — is the target of the evaluation that follows, and it motivates the regime-stratified analysis that carries our main result.

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026



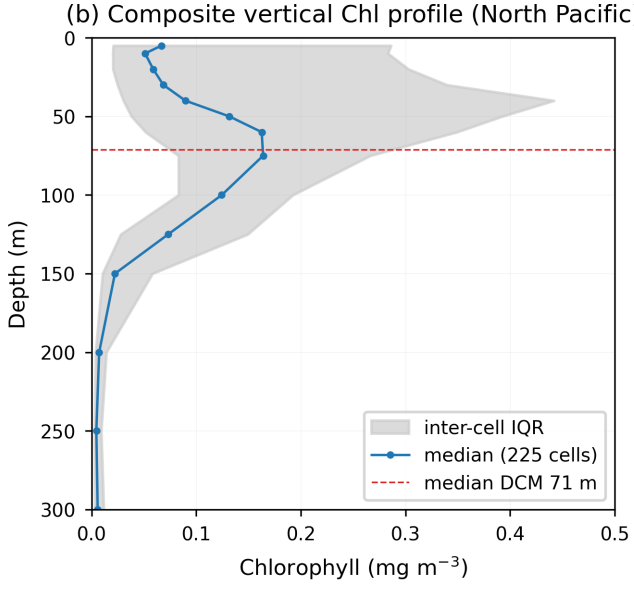


Figure 1. Observed BGC-Argo footprint and composite chlorophyll profile in the North Pacific. (a) Spatial footprint of BGC-Argo observed cells, coloured by the number of profiles per five-degree cell; land is masked in black. (b) Composite vertical chlorophyll profile over the full observed population (225 cells, 12,583 profiles): the solid line is the median of the reprocessed BGC-Argo chlorophyll and the shaded band is the inter-cell interquartile range. The dashed line marks the median observed DCM depth (71 m). The profile is descriptive and defines the observed vertical structure only.

3.2 Model biases in vertical chlorophyll structure are widespread and shared across the ensemble

Comparison of the CMIP6 models with the BGC-Argo diagnostics shows that biases in vertical chlorophyll structure are spatially widespread rather than confined to isolated cells (Fig. 2). Across the comparison cells, the

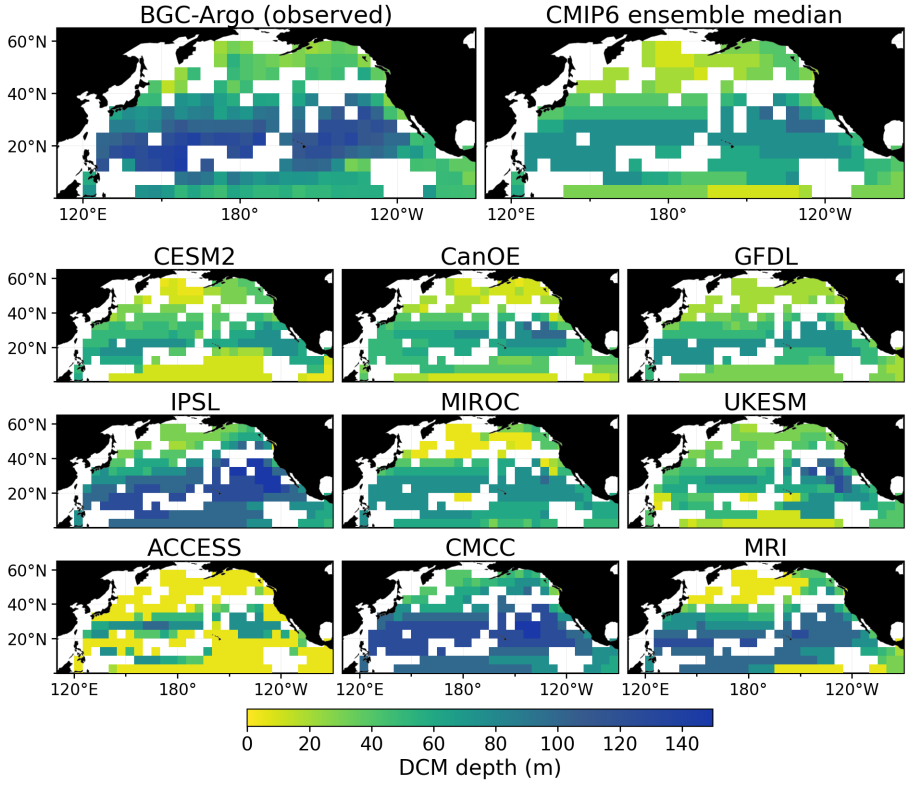
https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

models tend to place the chlorophyll maximum too shallow, to allocate a smaller fraction of chlorophyll to the subsurface, and to express a weaker vertical chlorophyll contrast (DCM prominence) than observed. These diagnostic biases appear across much of the sampled North Pacific and in multiple diagnostics simultaneously, indicating that they reflect a broad feature of the modelled vertical structure rather than a metric-specific artifact. We interpret these patterns as associations between modelled and observed structure and make no causal attribution here.

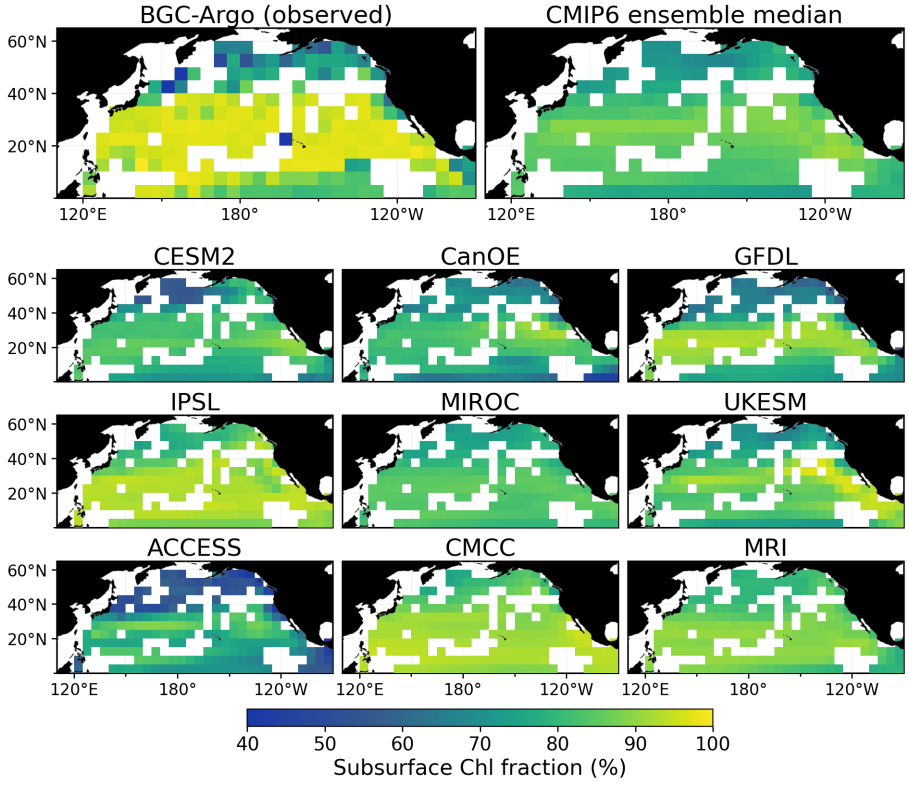
Stratifying the same comparison by model separates a component common to the ensemble from a model-dependent one (Fig. S2; 1,809 model–cell pairs across 201 cells, balanced at 201 per model). Two of the four diagnostics are one-signed across the ensemble: the median subsurface chlorophyll fraction bias is negative in every model (model-level range −2.0 to −17.9 percentage points), and the median DCM prominence ratio is below unity in every model (0.29 to 0.65). The under-allocation of subsurface chlorophyll and the weakening of vertical contrast are therefore ensemble-wide tendencies rather than the behaviour of a subset of models. The other two are not one-signed in the same way: the median surface chlorophyll ratio ranges from 1.10 to 4.77, and the median DCM-depth bias ranges from −44.4 to +20.9 m, varying in sign among models. We therefore treat the result as a systematic ensemble-wide tendency in subsurface allocation and vertical contrast, accompanied by model-to-model spread, and we do not rank the models. Because a given cell contributes to several model–cell pairs, the pooled distributions are interpreted descriptively and without significance tests.

Beyond this shared tendency, the per-model panels of Fig. 2 show that the spatial expression of the biases varies substantially among models — in location, in magnitude, and, for DCM depth, in sign — so the ensemble tendency is not a uniform shallow or weak bias reproduced by every model. How the models differ in the vertical structure they produce is examined next.

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026



https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026



https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

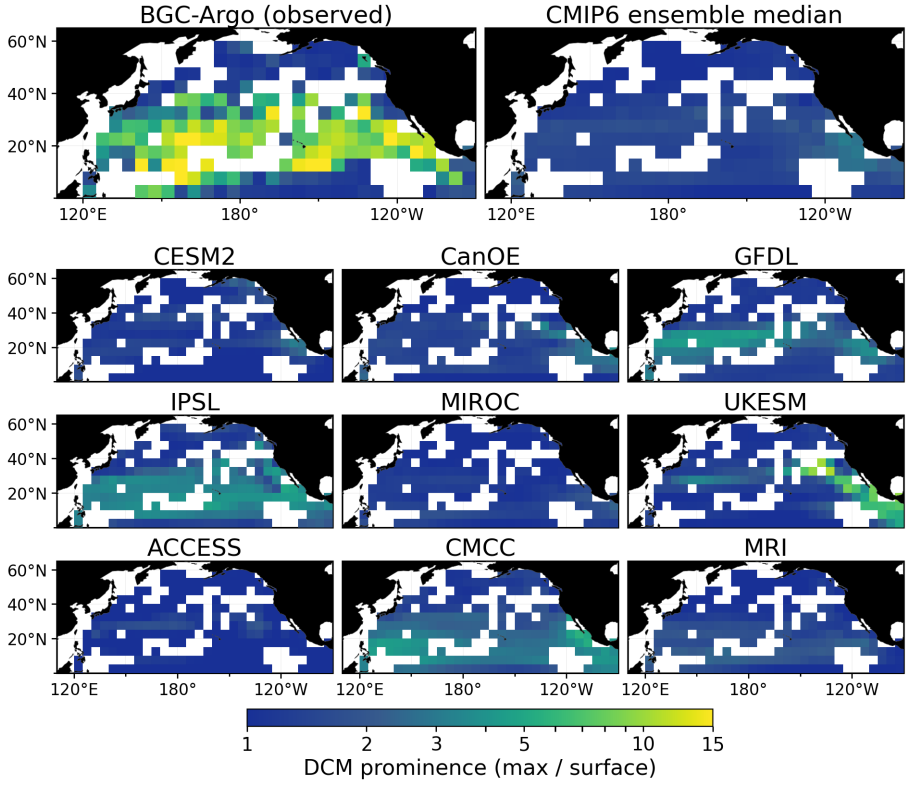


Figure 2. Spatial expression of the three vertical-structure diagnostics: (a) DCM depth (values, common colour scale), (b) subsurface chlorophyll fraction (values, 40–100 %), and (c) DCM prominence (values, log scale 1–15). Each panel shows BGC-Argo, the CMIP6 ensemble median, and the nine models on a scale common to that panel; black is land, white is a cell with no data. Per-model bias magnitudes are quantified in Fig. S2.

3.3 The models diverge in how they represent vertical chlorophyll structure

The ensemble-level summary of Section 3.2 does not show how individual models arrive at it. Figure 3 therefore presents, model by model, the vertical structure of chlorophyll normalized to its own 0–10 m median, computed over all 201 comparison cells and without regime stratification. Because every panel uses the same cells and the same axes, and because the BGC-Argo inter-cell interquartile range and median are repeated unchanged in every panel as a common background reference, differences between panels are differences between models. This comparison is intended to expose systematic differences among the models; the size of each model's bias relative to the observations is quantified regime by regime in Section 3.4 rather than read from this figure.

The models differ markedly in the depth, the strength, and the spatial spread of their subsurface chlorophyll structure. At one extreme, ACCESS expresses essentially no subsurface maximum: its normalized median profile

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

decreases almost monotonically below the surface layer, so that a chlorophyll maximum exists only in the formal sense that the profile attains a maximum somewhere. CESM2, CanOE, GFDL, UKESM, MIROC, and MRI express a modest maximum in the upper part of the water column, and only IPSL and CMCC develop a maximum deep enough and strong enough to reach the lower part of the observed inter-cell band. The models' own inter-cell bands are also visibly narrower than the observed band in most panels, but we do not quantify that contrast from Fig. 3, because the pooled observed band combines within-regime spatial variability with the spread between shallow and deep observed-DCM regimes. Stratifying by regime removes the between-regime component; within the deep regime the widest inter-cell bands are those of GFDL, IPSL, and CMCC, and all nine models nevertheless remain narrower than the observations (Section 4.5; Fig. S3c). CESM2's chlorophyll is resolved only to about 137.5 m on the common grid (Section 2.3), so its profile is truncated at depth rather than missing there.

Read together with the horizontal maps of Fig. 2, these panels indicate that the ensemble-wide biases of Section 3.2 are not produced by nine similar models, but by models that represent the DCM in qualitatively different ways. We describe these differences here and do not rank the models; they are summarized quantitatively in Section 3.5.

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

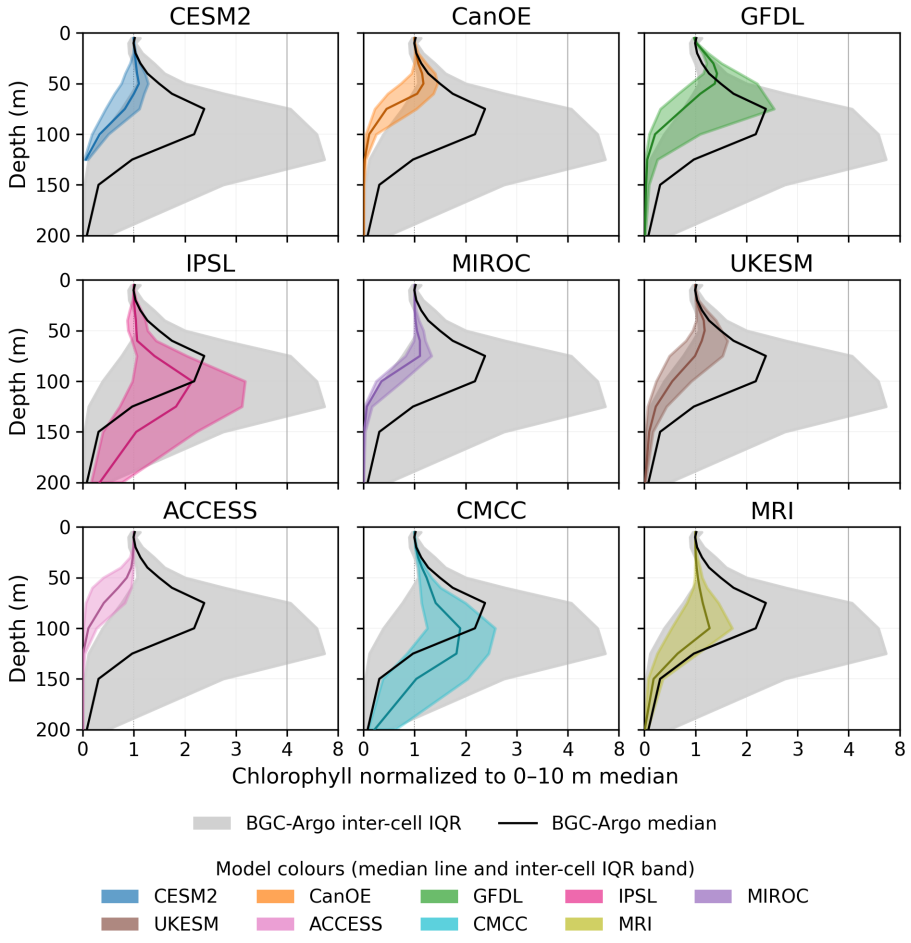


Figure 3. Per-model vertical chlorophyll structure across the North Pacific (all 201 comparison cells; not stratified by regime). The panels compare the models with one another against a common observed reference — the identical grey band (BGC-Argo inter-cell interquartile range) and black line (BGC-Argo median) in every panel — and are intended to reveal systematic differences among models, not to quantify each model's bias relative to the observations, which is done regime by regime in Fig. 4. In each panel the coloured band and line are that model's own inter-cell interquartile range and median of chlorophyll normalized to its 0–10 m median (median across cells; full names in Table 1). The x-axis is piecewise-linear (linear from 0 to 4, compressed above 4, tick at 8; the vertical line marks the break at 4) and the dotted line marks the surface reference (=1). CESM2's chlorophyll is resolved only to about 137.5 m on the common grid, so its profile is truncated at depth. The models differ markedly in the depth, strength, and spatial spread of their subsurface chlorophyll structure, and several (notably ACCESS) express little subsurface maximum at all. Pooled model–cell comparisons are not independent and are summarized descriptively without significance tests.

3.4 Within observed-DCM regimes, the models are too shallow and too weak, increasingly so where the observed DCM is deep

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

Section 3.1 showed that the observed DCM is a family of structures organized by depth. We therefore evaluate the models within observed-DCM regimes, stratifying all 201 comparison cells into shallow, intermediate, and deep regimes by tertiles of the observed DCM depth (67 cells each; thresholds 54.1 and 96.6 m; Section 2.5). Figure 4 shows the surface-normalized profiles within each regime, and Table 2 summarizes the regime-median diagnostics.

The observations progress systematically across the three regimes: the observed median normalized profile both deepens and sharpens. The peak of the plotted observed median profile rises from about 1.2 times the surface value near 40 m in the shallow regime, to about 3.3 times near 75 m in the intermediate regime, and to about 9.0 times near 125 m in the deep regime (Fig. 4). In the shallow regime the observed median profile is nearly flat, so the depth of its plotted peak is only weakly determined; the corresponding per-cell statistic is the more reliable one there. These plotted peak depths are the peaks of a median profile evaluated on the 14-level common depth grid and are a different statistic from the per-cell observed DCM depths of Section 3.1 (42.2, 71.1, and 112.1 m, obtained as the median of the per-cell profile maxima on the native grid); the two differ only through the vertical grid on which the maximum is located. The multipliers are likewise peaks of a normalized median profile, and are neither the per-cell prominence medians reported in Section 3.5 nor a substitute for them; they are sensitive to the surface normalization and are reported descriptively, with the corresponding magnitude claim carried by the prominence diagnostic (Table 2) and by Fig. 6.

The models do not follow the observations through this progression, and the mismatch grows as the observed maximum deepens. Across the shallow, intermediate, and deep regimes the regime-median DCM-depth bias is −9, −22, and −37 m, and the regime-median DCM prominence ratio is 0.89, 0.48, and 0.19 (Table 2): the models place the chlorophyll maximum increasingly too shallow, and express it increasingly too weakly, as the observed maximum deepens and sharpens. The shallow bias is not universal at the model level: within the deep regime CMCC and IPSL place their median maxima close to the observed depth, whereas the remaining models keep theirs in the upper part of the water column (Fig. 4). The subsurface chlorophyll fraction bias moves in the same direction, from essentially zero in the shallow regime to clearly negative in the intermediate and deep regimes (0.0, −10.2, and −9.2 percentage points), while the median model-to-observation surface chlorophyll ratio increases from 0.56 to 2.51 and 4.49 — that is, the ratio departs further from unity in precisely the regimes in which the models most under-represent the subsurface maximum. This ratio is the one diagnostic used here that depends on absolute chlorophyll magnitude; we therefore interpret its ordering across the regimes rather than the size of the departure, and we do not attribute that departure to the models rather than to the observations. The

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

models therefore do not merely mis-place a maximum of approximately the right shape: their representation of the observed structure degrades where the observed structure is most pronounced, and the deep observed-DCM regime accordingly provides the most stringent test of modelled vertical structure. Because the regimes are defined from the observed DCM depth, which also enters the DCM-depth bias, we present this regime dependence as an association and quantify the contribution of the stratification itself in Section 4.6. Following each model across the three regimes rather than comparing the models within a regime (Fig. S4, the transpose of Fig. 4), all nine models place their median maximum deeper in the deep regime than in the shallow regime, and in all nine the vertical contrast relative to the observations weakens as the regime deepens, with the largest shortfall in the deep regime. The pooled regime dependence of Table 2 is therefore expressed by every model rather than carried by a few.

The ensemble median (red in Fig. 4) is a poor summary of its own members in this respect. Because the models place their maxima at different depths, and mostly too shallow, the ensemble median profile is flatter than most of the individual model profiles and washes the maximum out, most visibly in the deep regime. We return to what this implies for ensemble-level evaluation in Section 4.2.

All distributions in this section pool model–cell pairs that share the same observation; they are summarized by medians and interquartile ranges, and no significance tests are applied.

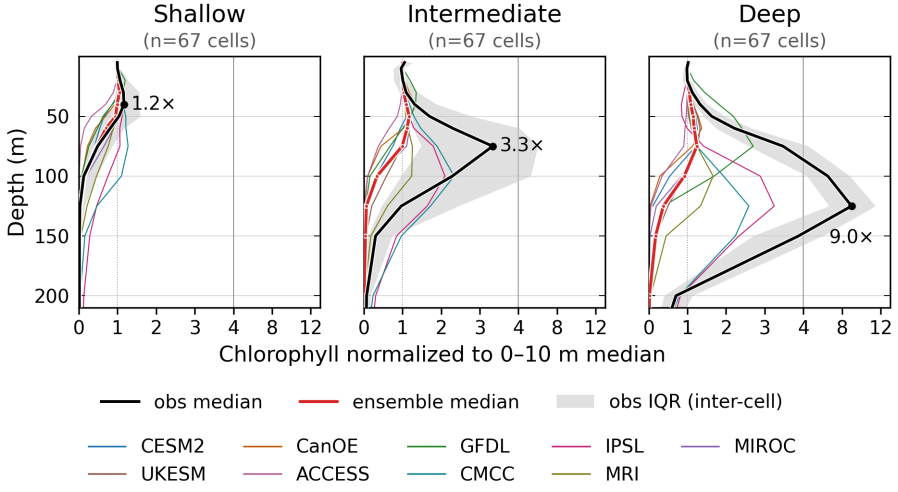


Figure 4. Regime-stratified profiles of chlorophyll normalized to the 0–10 m median, for observations (black, median across cells; grey band, inter-cell interquartile range), the nine CMIP6 models (coloured, per-model median across cells; see Table 1 for full names), and the ensemble (red, median of the nine model medians). Comparison cells are stratified into shallow, intermediate, and deep regimes by tertiles of observed DCM depth (n = 67, 67, and 67 comparison cells; all 201 comparison cells). The x-axis is piecewise-linear (linear

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

from 0 to 4, compressed above 4, with ticks at 8 and 12; the solid vertical line marks the break at 4) to resolve the crowded near-surface range where the models lie while retaining the observed deep maximum; the dotted line marks the surface reference (=1). Black dots and labelled multipliers give the peak value of the plotted observed median profile (1.2×, 3.3×, and 9.0× at 40, 75, and 125 m on the 14-level common grid); these are normalization-sensitive and are corroborated by the integrated prominence diagnostic (Table 2) and Fig. 6. The observed DCM-depth medians (42.2, 71.1, and 112.1 m; per-cell argmax on the native grid) differ from these plotted peak depths owing to grid resolution, and in the shallow regime also because the observed median profile is nearly flat, so that the depth of its plotted peak is weakly determined. Pooled model–cell pairs are not independent; distributions are summarized descriptively without significance tests.

Table 2. Regime-median vertical-chlorophyll diagnostics. All 201 comparison cells, stratified into observed-DCM regimes by tertiles of the observed DCM depth (67 cells and 603 model–cell pairs per regime). The observed DCM depth is the median over the cells of each regime of the per-cell profile maximum on the native grid; the modelled DCM depth is the median over the pooled model–cell pairs of each regime of the per-pair profile maximum on the common depth grid; the biases and the ratio are medians over the same pooled pairs of the per-pair difference or ratio. Each row is therefore an independently computed median, and the rows are not related by arithmetic: a median of differences is not the difference of medians. Pooled pairs are not independent and no significance tests are applied.

	Shallow (67 cells)	**Intermediate (67 cells)**	**Deep (67 cells)**
Observed DCM depth (m; per-cell argmax, native grid)	42.2	71.1	112.1
Modelled DCM depth (m; per-pair argmax, common grid)	30	50	75
Surface chlorophyll ratio (model / BGC-Argo)	0.56	2.51	4.49
DCM-depth bias (model − BGC-Argo; m)	−9	−22	−37
Subsurface chlorophyll fraction bias (model − BGC-Argo; percentage points)	0.0	−10.2	−9.2
DCM prominence ratio (model / BGC-Argo)	0.89	0.48	0.19

3.5 A typology of DCM representation

The per-model profiles of Section 3.3 suggest that the models differ in kind and not only in degree. Figure 5 summarizes each model by two quantities computed over the 201 comparison cells: the median depth of its chlorophyll maximum, and the median prominence of that maximum, defined for each field as its own maximum chlorophyll over 5–145 m divided by its own 0–10 m surface chlorophyll. This prominence is each field's own vertical contrast; it is a different quantity from the model-to-observation prominence ratio reported in Sections 3.2, 3.4, and 3.6, and the two are not compared with one another.

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

Three groups separate. ACCESS has a median prominence of 1.00: its profile is essentially flat, and the depth at which its maximum is found (5 m) therefore carries no information about a subsurface feature — an illustration that DCM depth alone can misclassify a profile that has no subsurface maximum to locate. CESM2, CanOE, GFDL, UKESM, MIROC, and MRI form a shallow group, with median maxima at 40–75 m and prominences of 1.17–1.51. CMCC and IPSL form a deep group, with median maxima at 100 m and prominences of 2.38 and 2.89. The observations lie outside all three groups: the observed median DCM depth is 71 m and the observed median prominence is 3.6, so even the deep group, which comes closest, does not reach the observed combination of depth and vertical contrast. The observed depth here is the median of the per-cell maxima on the native grid, whereas the model depths are located on the common depth grid; the two differ only by grid resolution.

These groups are descriptive. We draw no boundaries in the figure, we do not rank the models, and we do not attribute the grouping to particular model formulations; possible reasons are considered as hypotheses in Section 4.3.

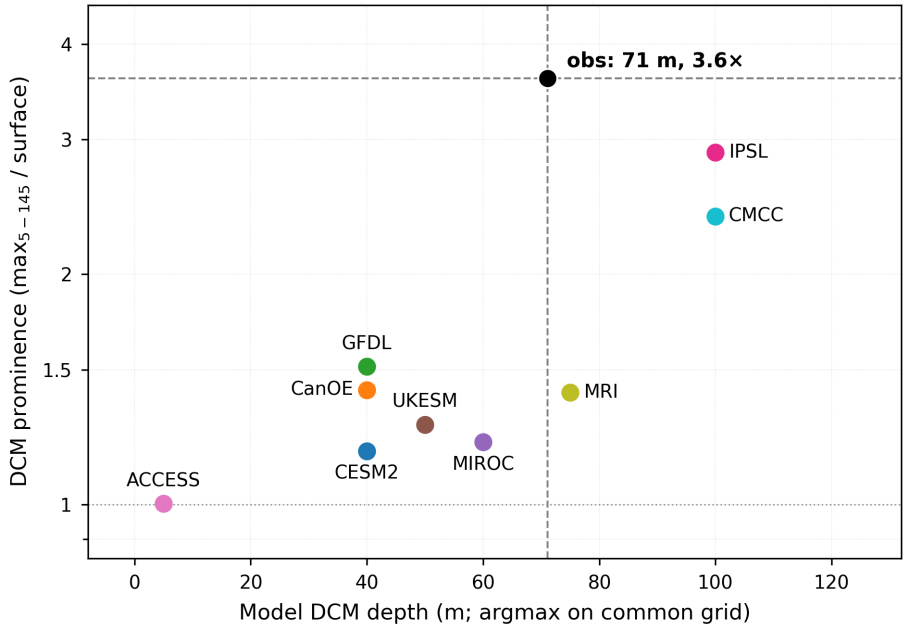


Figure 5. Model typology of deep chlorophyll maximum (DCM) representation. Each point is one CMIP6 model, positioned by its median DCM depth (x; depth of the chlorophyll maximum by argmax on the common 14-level grid) and its median DCM prominence (y, logarithmic; the model's own maximum chlorophyll over 5–145 m divided by its 0–10 m surface chlorophyll), each taken as the per-model median across the 201 comparison cells (see Table 1 for full names). The black point marks the observations (per-cell median across the 201 cells; observed DCM depth 71 m on the native grid, observed prominence 3.6), and the dashed crosshair marks these observed

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

values; the dotted line at prominence = 1 marks the absence of a subsurface maximum. This prominence is each field's own surface-to-maximum contrast and is distinct from the model-to-observation prominence ratio, which is below unity in every model over the full comparison (Fig. S2) and is reported by regime in Table 2 and within the surface-matched band in Fig. 6. The models separate into flat (ACCESS), shallow, and deep (CMCC, IPSL) DCM-representation groups, and none reaches the observed depth-and-prominence envelope; ACCESS returns a shallow argmax depth but a prominence near unity — an essentially flat profile — illustrating that DCM depth alone can misclassify a profile that lacks a subsurface maximum.

3.6 Surface chlorophyll agreement does not constrain subsurface allocation, DCM prominence, or DCM depth

The preceding sections evaluate the models over the full comparison, in which modelled and observed surface chlorophyll generally differ. A natural objection is that the vertical-structure biases simply follow from that surface mismatch. To test this, we restricted the comparison to the surface-matched band, defined as model–cell pairs with |log2(model/BGC-Argo surface chlorophyll ratio)| < 0.5, and quantified the vertical-structure biases diagnostic by diagnostic (Fig. 6; 263 model–cell pairs across 102 cells, drawn from the source population of 1,809 pairs across 201 cells).

Within this band, the subsurface chlorophyll fraction bias is systematically negative, with a pair-based median of −5.5 percentage points (interquartile range −10.1 to −1.2) and a cell-based median of −5.6 percentage points (interquartile range −9.1 to −2.9); the interquartile range lies entirely below zero on both bases (Fig. 6b). DCM prominence within the band is likewise below unity, with a pair-based median ratio of 0.81 and a cell-based median of 0.79, indicating a weaker vertical chlorophyll contrast than observed even where surface chlorophyll agrees (Fig. 6c). Because DCM prominence is defined with surface chlorophyll in its denominator, we interpret only its offset within the band and not its slope across the band. The DCM-depth bias within the band is a modest shallow bias with substantial residual scatter: the pair-based median is −17 m (interquartile range −33.9 to +8.3 m) and the cell-based median is −14 m (interquartile range −25.4 to +3.0 m), so the interquartile range spans zero on both bases and surface agreement does not correspond to a single systematic depth bias (Fig. 6a). Within the band, the sign of the per-model median DCM-depth bias is not consistent across models — shallow in five of the nine models and deep in the other four (Fig. 6a). This band-level behaviour of DCM depth differs from the regime-stratified result of Section 3.4; we reconcile the two in Section 4.1.

Taken together, these results indicate that surface chlorophyll agreement is not sufficient to constrain the vertical allocation of chlorophyll, its vertical contrast, or the depth of the chlorophyll maximum. Because a single cell can

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

contribute to several model–cell pairs, the pooled distributions are reported descriptively, on a pair and a cell basis in parallel, and without significance tests.

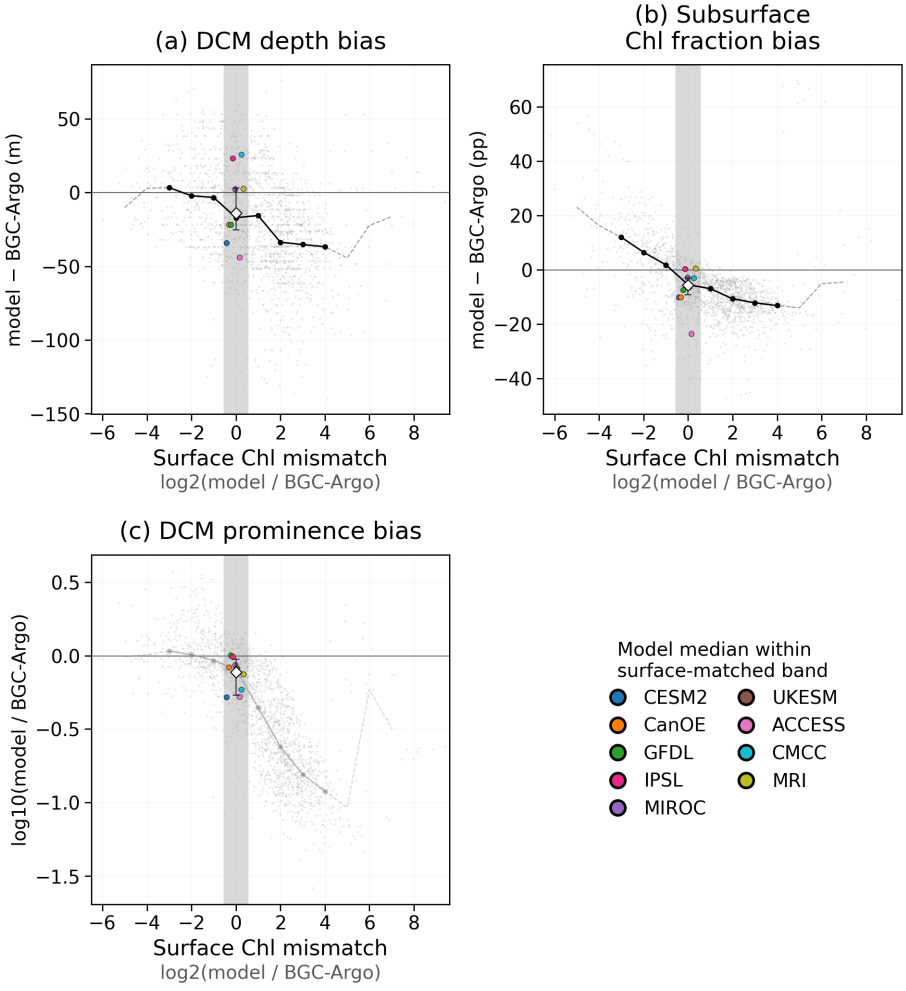


Figure 6. Vertical-structure biases within the surface-matched band as a function of surface chlorophyll mismatch (263 pairs across 102 cells; source population 1,809 pairs across 201 cells). (a) DCM-depth bias, (b) subsurface chlorophyll fraction bias, (c) DCM prominence (log10 model/BGC-Argo), versus the surface chlorophyll ratio. Grey points: individual pairs; black line: pooled binned median (solid where bins contain ≥40 pairs, grey-dashed where sparse, n<40); grey band: surface-matched subset (|log2 ratio|<0.5); white diamond: band cell-based median ± IQR; coloured points: per-model band medians (full names in Table 1). No significance tests are applied because pooled pairs are not independent; the sparse region (x≥5) is not interpreted. In panel (c), the surface-mismatch gradient is drawn faintly and is not interpreted: because DCM prominence is normalised by the surface value, a gradient against surface mismatch is in part definitional; prominence is interpreted by its band offset only.

3.7 The subsurface offset is independent of the BGC-Argo chlorophyll reprocessing

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

Because the observational reference uses a reprocessed BGC-Argo chlorophyll (Section 2.2), we tested whether the surface-matched subsurface offset depends on this choice. The reprocessing rescales chlorophyll by a single, depth-independent factor per cell. The subsurface chlorophyll fraction, the DCM prominence, and the DCM depth are ratios or extrema of the profile and are therefore mathematically invariant to such a per-cell rescaling; only the surface chlorophyll — and hence surface-matched band membership — changes between the reprocessed and the earlier factor-two definitions. Recomputing the surface-matched band under the factor-two definition therefore changes which cells satisfy the surface-matching criterion (263 pairs / 102 cells for the reprocessed definition; 218 / 89 for the factor-two definition; 151 / 70 for their intersection), but the subsurface chlorophyll fraction bias remains negative under all three definitions (pair-based medians −5.5, −4.7, and −4.5 percentage points), with the upper quartile no higher than +0.04 percentage points — that is, at most marginally above zero — in every case. The negative subsurface offset is therefore a property of the vertical allocation itself, not of the reprocessing (Fig. S1).

4 Discussion

4.1 Reconciling the regime-stratified and surface-matched views of the depth bias

The DCM-depth bias behaves differently in the two comparisons presented above, and the difference is informative rather than contradictory. Across observed-DCM regimes the bias is negative in every regime and becomes steadily more negative as the observed maximum deepens (−9, −22, and −37 m; Table 2). Within the surface-matched band, by contrast, the pooled bias is only modestly shallow, its interquartile range spans zero, and the sign of the per-model median is shallow in five of the nine models and deep in the other four (Section 3.6).

The two views are reconciled by the composition of the surface-matched band. Surface matching selects cells in which modelled and observed surface chlorophyll agree, and this selection is not neutral with respect to the observed DCM depth: of the 102 cells that enter the surface-matched comparison, 48 come from the 67 shallow-regime cells and 40 from the 67 intermediate-regime cells, but only 14 from the 67 deep-regime cells. The surface-matched band is therefore strongly depleted in exactly the regime in which the shallow bias is largest, and its pooled DCM-depth statistics are dominated by regimes in which that bias is small. The apparent absence of a systematic depth bias under surface matching is thus a property of the sampled subset rather than evidence that the depth bias is unsystematic.

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

This depletion has a practical consequence. Conditioning on surface agreement does not merely fail to constrain vertical structure (Section 3.6); it also preferentially removes the water columns whose vertical structure is most demanding to reproduce. The surface-matched result should therefore be read as a conservative test: it is obtained on the subset least likely to expose vertical-allocation error, and the subsurface-fraction bias is negative there nonetheless, with an interquartile range entirely below zero. Stratifying this surface-matched subsurface-fraction bias by observed-DCM regime, and repeating the surface match under an absolute criterion (|model − observed surface chlorophyll| < 0.05 mg m $^{-3}$; the threshold is nominal — the negative sign is retained at 0.05, 0.10 and 0.20 mg m $^{-3}$, although the magnitude is not, Fig. S5) in addition to the ratio criterion used above, shows the negative bias in every regime and under both criteria, so the subsurface deficit is not an artefact of the surface-match definition (Fig. S5); its magnitude and the ordering among regimes do depend on the definition, and under the stringent ratio criterion the deep regime is represented by only a small, biased subset (40 pairs in 14 cells), consistent with the depletion noted above.

4.2 Ensemble-level agreement in DCM depth is not evidence of per-model fidelity

The vertical-structure biases differ in how consistent they are across models, and this difference shapes how ensemble-level agreement should be read. The subsurface chlorophyll fraction bias and the DCM prominence ratio are one-signed across the ensemble: over the full comparison the fraction bias is negative and the prominence ratio below unity in every model (Fig. S2). The DCM-depth bias is not: over the full comparison its per-model median is shallow in seven of the nine models and deep in the other two, spanning −44.4 to +20.9 m (Fig. S2), and within regimes the models separate into those that keep their maximum in the upper water column and the two — CMCC and IPSL — that reach the observed depth in the deep regime (Fig. 4).

The spread in depth has two distinct consequences for ensemble-level summaries. First, because the models place their maxima at different depths, the ensemble median profile is flatter than most of the individual model profiles and washes the maximum out; Figure 4 makes this explicit within each regime. This effect requires only dispersion in depth, and not errors of opposing sign. Second, where too-shallow and too-deep models offset one another, an ensemble-level summary of DCM depth can sit closer to the observations than most of its members do. Apparent ensemble agreement in DCM depth is therefore better read as cancellation between opposing per-model errors than as per-model fidelity, and it is not evidence that the individual models represent the DCM. Fraction and prominence, being consistently one-signed, are not subject to this cancellation and are the more reliable ensemble-level indicators of vertical-structure bias. We therefore treat the subsurface fraction and

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

prominence biases as the primary ensemble-level evidence, and interpret ensemble DCM-depth agreement cautiously. These patterns are diagnostic; we do not attribute them to specific mechanisms.

4.3 What the model typology implies

The models do not differ only in the magnitude of a shared bias; they differ in kind. Section 3.5 separates them into three types of DCM representation: a flat type that expresses essentially no subsurface maximum (ACCESS), a shallow type with a modest maximum in the upper water column (CESM2, CanOE, GFDL, UKESM, MIROC, and MRI), and a deep type with a maximum near 100 m and a markedly stronger vertical contrast (CMCC and IPSL). None of the three reaches the observed combination of DCM depth and prominence.

Two consequences follow for evaluation practice. First, a single diagnostic can misclassify a model: ACCESS returns a DCM depth in the same way as any other model, but its prominence of 1.00 shows that there is no subsurface maximum for that depth to describe. Depth and vertical contrast have to be read together, which is why we report them jointly rather than reducing either to a skill score. Second, because the three types are qualitatively different, an ensemble-mean chlorophyll profile is not a compromise among similar structures but a blend of structurally different ones, which reinforces the reading of Section 4.2.

Why the models divide in this way is beyond the diagnostics used here. Observed DCM depth and intensity co-vary with light availability, nutrient supply, and mixing (Cornec et al., 2021; Barbieux et al., 2019), so it is plausible that the differences reflect how each model represents the vertical structure of light attenuation, the nutricline, and vertical mixing, together with the phytoplankton physiology — including photoacclimation — that converts biomass into chlorophyll. We offer this as a hypothesis to be tested against the corresponding model fields, and not as an attribution supported by the present analysis.

4.4 Relation to previous evaluations of marine biogeochemical models

Previous CMIP-generation evaluations have relied substantially on surface, upper-ocean, and vertically integrated diagnostics, including satellite-based surface chlorophyll (Séférian et al., 2020; Kwiatkowski et al., 2020). More recent work has begun to evaluate DCM-related and vertical chlorophyll diagnostics in specific regions, notably across CMIP6 models in the Southern Ocean (Cheng et al., 2025), and observational syntheses have mapped observed subsurface chlorophyll structure globally, including the North Pacific (Yasunaka et al., 2022). Our contribution differs in what is evaluated and how. Rather than characterizing a mean vertical bias, we evaluate the models against the range of observed structures: we stratify by the observed DCM depth, so that the models

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

are tested where the observed maximum is weakly and where it is strongly expressed, and we characterize each model by the kind of DCM it produces rather than by a skill score. The result is a regime-dependent picture — fidelity degrades where the observed DCM is deepest and most sharply expressed — together with a typology rather than a ranking. The surface-matched comparison then complements observational analyses showing that surface chlorophyll alone does not uniquely determine subsurface or column chlorophyll structure (Cornec et al., 2021; Tanner et al., 2024), and extends that non-uniqueness from an observational property to a concrete limitation of surface-based model evaluation.

4.5 Implications for model evaluation and development

Because surface chlorophyll agreement does not guarantee subsurface fidelity, model skill assessed against surface chlorophyll alone may overstate the fidelity of the simulated chlorophyll field, and improvements that reduce surface chlorophyll bias need not reduce the vertical-structure bias. The regime dependence sharpens this point: the model-to-observation surface chlorophyll ratio is largest in the deep regime (median 4.49; Table 2), which is also the regime in which they most under-represent the subsurface maximum, so tuning toward surface chlorophyll in these regions would act on the diagnostic that is least informative about the vertical-structure error.

The models also under-represent the spatial diversity of vertical structure, and they do so in a way that is not a simple attenuation. Within each observed-DCM regime we compare the inter-cell interquartile spread of the surface-normalized profiles between each model and the observations, integrated over 40–150 m (Section 2.4; Fig. S3). Within the intermediate and the deep regimes, every one of the nine models produces a narrower inter-cell spread than observed; in the deep regime the per-model spread is between about 5 % and 26 % of the observed spread (Fig. S3b, c). The shallow regime behaves differently, and instructively: there the observed profile is close to flat — its median normalized profile peaks at only about 1.2 times the surface value (Section 3.4) — and CMCC and IPSL produce more inter-cell spread than observed (Fig. S3a). We do not quantify this excess in the shallow regime. The shallow regime is not a single environment: its cells fall into two groups, equatorward and poleward of about 25° N (Fig. A1), whose observed inter-cell spreads differ by roughly a factor of five, and the pooled spread of the regime is close to the narrower of the two. The excess is present against either group, but its size depends on which cells are pooled. These are the two models that place their chlorophyll maximum deepest overall (Section 3.5), and their excess spread is consistent with their imposing a subsurface maximum in cells where the observations show little vertical structure. The ensemble's error is therefore not a simple attenuation of the observed variability: where the observed maximum is strongly expressed every model produces less spatial diversity than observed, whereas where the observed profile is nearly flat the deep-type

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

models produce more. The comparison is made within regimes, so the between-regime component of the observed spread is excluded (Section 3.3), and it is made on the same normalized profiles that Fig. S3 displays, so that the contrast refers to the differences that the figure resolves rather than to relative differences among near-zero values. The ratios are reported descriptively, without significance tests, because the cells contributing to a spread are not independent.

Depth-resolved constraints, such as those now available from BGC-Argo, are therefore useful for diagnosing and targeting the vertical distribution of chlorophyll. We emphasize that our diagnostics are relative and structural and are thus robust to the well-known uncertainty in absolute chlorophyll magnitude; we do not interpret absolute chlorophyll differences. We also stress that this analysis is not a ranking of models: the biases have a component common to all nine models and a model-dependent component, and we treat the ensemble accordingly.

4.6 Limitations

Several limitations bound the interpretation. First, the observed BGC-Argo years and the CMIP6 historical period do not coincide; we frame the comparison as a structural, climatological assessment. Because the diagnosed structural differences are large — factors of up to about five in surface chlorophyll and DCM-depth differences of tens of metres — they are large relative to the decadal changes in upper-ocean chlorophyll structure expected between the two windows, so the temporal offset is unlikely to account for the systematic vertical-structure biases.

Second, the BGC-Argo footprint samples the North Pacific unevenly; our results describe the BGC-Argo-sampled North Pacific and are not an area-weighted representation of the full basin. The subsurface-fraction diagnostic is defined over the 7.5–137.5 m window resolved by all nine models; extending the window to 175 m for the eight models that resolve it leaves the fraction essentially unchanged and does not affect the model-minus-observation bias (Supplement).

Third, BGC-Argo chlorophyll is fluorescence-based and subject to calibration and near-surface quenching; we use the adjusted, community-reprocessed product and rely on relative, structural diagnostics, and we show that the surface-matched subsurface offset is retained under both chlorophyll definitions (Fig. S1). A chlorophyll maximum also need not coincide with a phytoplankton biomass maximum, because photoacclimation can raise chlorophyll-to-carbon ratios at depth (Cornec et al., 2021); our diagnostics therefore describe the vertical allocation of chlorophyll rather than of biomass. Consistent with this, a companion analysis of the same North

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

Pacific BGC-Argo profiles finds the chlorophyll DCM frequently displaced from the particle-backscattering maximum (Hashioka and Aita, 2026a).

Fourth, the regimes are defined from the observed DCM depth of each comparison cell, and the DCM-depth bias is evaluated as model minus observation within those regimes. Stratifying on an observed quantity that also enters the bias means that sampling error in the cell-median observed DCM depth will tend to steepen the apparent regime dependence of that bias, because the cells assigned to the deep regime are, on average, those whose observed DCM depth is estimated on the high side. We therefore quantified this effect rather than only noting it. The cell medians are formed from a median of 44 profiles per cell (12,500 profiles over the 201 comparison cells), and the reliability of the cell-median observed DCM depth — the fraction of its between-cell variance that is not sampling error — is estimated at 0.96 both from the dependence of the between-cell variance on the number of profiles per cell and from a variogram-based estimate. The corresponding regression-to-the-mean displacement of the regime-mean observed depth is about +1 m in the shallow regime and about −1.5 m in the deep regime, which accounts for about 9 % of the −28 m difference in median DCM-depth bias between the shallow and the deep regimes; the remaining 91 % is not attributable to the stratification. Two further observations point the same way. The modelled DCM depth deepens across the regimes rather than remaining fixed (30, 50, and 75 m; Table 2), spanning about two-thirds of the observed deepening, so the models follow the observed change only partially rather than not at all. This aggregate slope masks a wide spread among the models: computed model by model, the deepening of each model's regime-median DCM depth relative to the observed deepening ranges from about 0.3 to about 1.0 (Fig. S4), so that individual models span from tracking roughly a third of the observed deepening to matching it, with the ensemble value near the two-thirds given above. We do not report the individual per-model slopes, because the modelled DCM depth is quantized on the common depth grid, so that the difference between adjacent models corresponds to about one grid step; no significance is attached to the spread. And the regime dependence is not confined to the stratifying variable: the prominence ratio and the subsurface-fraction bias, neither of which is used to define the regimes, show the same dependence — although the observed prominence and the observed DCM depth are themselves related, so this is corroboration rather than an independent test. We therefore report the regime dependence of the DCM-depth bias as a measured association whose steepness is only slightly inflated by the stratification, rather than as a calibrated slope, and note that a stratification derived independently of the evaluated observations — for example from a split of the profiles within each cell — would provide a stronger test.

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

Fifth, the normalized profiles of Figs. 3 and 4 are ratios to the surface value and their magnitude is correspondingly sensitive to it; quantitative statements about the strength of the chlorophyll maximum are carried by the prominence diagnostic and by Fig. 6 rather than by the appearance of the normalized profiles. The observed inter-cell spread used in Section 4.5 likewise contains a sampling-noise component. Because this noise inflates the observed spread, which is the denominator of the ratios reported there, those ratios are conservative: removing the noise contribution would raise them, strengthening the excess found in the shallow regime and leaving unchanged the finding that the models are narrower than observed in the intermediate and the deep regimes.

Finally, because a single cell can contribute to several model–cell pairs, the pooled distributions are not independent; we report them descriptively, without significance tests, and separately on a pair and a cell basis.

4.7 Future work

This analysis is diagnostic rather than mechanistic. A natural next step is to relate the diagnosed vertical-structure biases, and the typology of Section 3.5 in particular, to their potential drivers — biases in stratification and mixed-layer depth, nutricline structure, or light — using variables available from the same models and, where sample size permits, from BGC-Argo; the grouping of the models into flat, shallow, and deep types offers a natural set of contrasts for such an analysis. A stratification derived independently of the evaluated observations (Section 4.6) would strengthen the regime-dependent result. Extending the evaluation to other basins would establish whether the North Pacific result generalizes, and combining BGC-Argo with satellite phytoplankton size-class or backscattering information could further constrain the vertical structures that surface chlorophyll cannot resolve. A companion analysis takes this step for float-based particle backscattering, mapping the displacement between the chlorophyll DCM and the backscattering maximum across the North Pacific (Hashioka and Aita, 2026a).

5 Conclusions

Using BGC-Argo chlorophyll profiles over the North Pacific, we evaluated the vertical chlorophyll structure of nine CMIP6 biogeochemical models against the range of vertical structures that is actually observed. The observations show that the deep chlorophyll maximum is not a single feature but a family of structures that deepens and sharpens together across the basin. The models represent this family poorly and in a regime-dependent way: they allocate a smaller fraction of chlorophyll to the subsurface and express a weaker vertical chlorophyll contrast than observed — over the full comparison the subsurface-fraction bias is negative and the

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

prominence ratio below unity in all nine models — and they place the chlorophyll maximum increasingly too shallow as the observed maximum deepens. They also differ in kind rather than only in degree, separating into flat, shallow, and deep types of DCM representation, none of which reaches the observed combination of depth and prominence; because they place their maxima at different depths, the ensemble median washes the maximum out and is a poor summary of its own members, so ensemble-level agreement in DCM depth would reflect cancellation among per-model errors of opposing sign rather than fidelity in any individual model. These vertical-allocation biases are not diagnosable from the surface: they persist within the surface-matched subset, where modelled and observed surface chlorophyll already agree and where the subsurface chlorophyll fraction bias is negative with an interquartile range entirely below zero. We conclude that surface chlorophyll agreement is not sufficient to diagnose biases in the vertical allocation of chlorophyll, and that surface-focused evaluation can leave systematic subsurface structural biases undetected. In this sense the surface-matched comparison converts a known observational non-uniqueness — that surface chlorophyll does not uniquely determine subsurface structure — into a concrete limitation of surface-based model evaluation. It also follows that improvements which reduce surface chlorophyll bias need not, by themselves, reduce the vertical-structure bias. The compact, reproducible set of diagnostics used here — surface chlorophyll, DCM depth, subsurface chlorophyll fraction, and DCM prominence, evaluated within observed-DCM regimes and under explicit surface matching — can be applied to other regions and model ensembles.

Appendix A: Observed DCM-depth regimes

Comparison cells were grouped into shallow, intermediate, and deep observed-DCM regimes by tertiles of their observed DCM depth (Section 2.5). The tertile thresholds, 54.1 and 96.6 m, are the lower and upper terciles of the observed DCM depth across the 201 comparison cells and divide them into three groups of 67 cells, with median observed DCM depths of 42.2, 71.1, and 112.1 m. The spatial distribution of the regimes is shown in Fig. A1; the regime-median diagnostics are given in Table 2 and the regime-stratified profiles in Fig. 4.

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

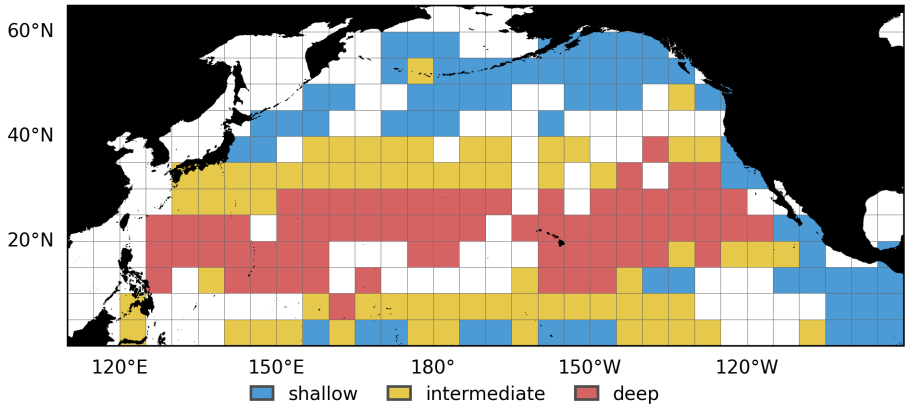


Figure A1. Observed BGC-Argo chlorophyll DCM-depth regime groups over the North Pacific. Each occupied 5° comparison cell (201 cells) is coloured by its observed DCM-depth tertile group (shallow, intermediate, deep), defined by tertiles of observed DCM depth across the 201 comparison cells (thresholds 54.1 and 96.6 m; 67 cells each; median observed DCM depths 42.2, 71.1, and 112.1 m). Land is masked in black; white cells are occupied by observations but do not enter the model–cell comparison. The regime grouping is based on observed DCM depth only (not on any model field), and the colour scale is categorical. Model biases stratified by these regimes are shown in Table 2 and Fig. 4.

Code and data availability

The analysis-ready data, the code that generates them, and the figure-generating scripts are archived at Zenodo (https://doi.org/10.5281/zenodo.20997630) under the Creative Commons Attribution 4.0 licence (Hashioka and Aita, 2026b). Raw BGC-Argo, CMIP6, satellite, and map source data are not redistributed there and are available from their original archives: BGC-Argo profiles from the Argo Global Data Assembly Centre; CMIP6 ocean chlorophyll (variable chl, table Omon, experiment historical) from the Earth System Grid Federation for the nine source_id/variant_label/grid_label combinations listed in Table 1; and the SEANOE physiological look-up table (Sauzède et al., 2026; https://doi.org/10.17882/105732). The factor-two chlorophyll reference is Roesler et al. (2017).

Author contributions

T.H. designed the study, processed the BGC-Argo and CMIP6 chlorophyll datasets, developed the diagnostic framework, performed the model–observation comparisons, created the figures, interpreted the results, and wrote the original manuscript draft. M.N.A. contributed to the scientific framing, interpretation of the North Pacific biogeochemical context, funding acquisition, and manuscript review and editing. Both authors discussed the results and approved the final manuscript.

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors gratefully acknowledge the International Argo Program and the national programs that contribute to it for collecting and making BGC-Argo data publicly available. The authors also acknowledge the World Climate Research Programme, which coordinated and promoted CMIP6, the climate modelling groups (listed in Table 1) for producing and making available their model output, and the Earth System Grid Federation for archiving and distributing it.

Financial support

This study was supported by institutional funding from the Japan Agency for Marine-Earth Science and Technology (JAMSTEC) and by the Japan Science and Technology Agency (JST) CREST program through the project "Carbon cycle and Lower trophic level oceanic organisms Investigation for Future Environments (C-LIFE)" (grant no. JPMJCR24J1).

References

Barbieux, M., Uitz, J., Gentili, B., Pasqueron de Fommervault, O., Mignot, A., Poteau, A., Schmechtig, C., Taillandier, V., Leymarie, E., Penkerc'h, C., D'Ortenzio, F., Claustre, H., and Bricaud, A.: Bio-optical characterization of subsurface chlorophyll maxima in the Mediterranean Sea from a Biogeochemical-Argo float database, Biogeosciences, 16, 1321-1342, https://doi.org/10.5194/bg-16-1321-2019, 2019.

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., Bekki, S., Bonnet, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Caubel, A., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., D'Andrea, F., Davini, P., de Lavergne, C., Denvil, S., Deshayes, J., Devilliers, M., Ducharne, A., Dufresne, J.-L., Dupont, E., Éthé, C., Fairhead, L., Falletti, L., Flavoni, S., Foujols, M.-A., Gardoll, S., Gastineau, G., Ghattas, J., Grandpeix, J.-Y., Guenet, B., Guez, L. E., Guilyardi, E., Guimberteau, M., Hauglustaine, D., Hourdin, F., Idelkadi, A., Joussaume, S., Kageyama, M., Khodri, M., Krinner, G., Lebas, N., Levavasseur, G., Lévy, C., Li, L., Lott, F., Lurton, T., Luyssaert, S., Madec, G., Madeleine, J.-B., Maignan, F., Marchand, M., Marti, O., Mellul, L., Meurdesoif, Y., Mignot, J., Musat, I., Ottlé, C., Peylin, P., Planton, Y., Polcher, J., Rio, C., Rochetin, N., Rousset, C., Sepulchre, P., Sima, A., Swingedouw, D., Thiéblemont, R., Traore, A. K.,

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

Vancoppenolle, M., Vial, J., Vialard, J., Viovy, N., and Vuichard, N.: Presentation and Evaluation of the IPSL-CM6A-LR Climate Model, Journal of Advances in Modeling Earth Systems, 12, e2019MS002010, https://doi.org/10.1029/2019MS002010, 2020.

Cheng, M., Maher, N., and Ellwood, M. J.: Evaluating the performance of CMIP6 models in simulating Southern Ocean biogeochemistry, Biogeosciences, 22, 7269-7291, https://doi.org/10.5194/bg-22-7269-2025, 2025.

Christian, J. R., Denman, K. L., Hayashida, H., Holdsworth, A. M., Lee, W. G., Riche, O. G. J., Shao, A. E., Steiner, N., and Swart, N. C.: Ocean biogeochemistry in the Canadian Earth System Model version 5.0.3: CanESM5 and CanESM5-CanOE, Geoscientific Model Development, 15, 4393-4424, https://doi.org/10.5194/gmd-15-4393-2022, 2022.

Cornec, M., Claustre, H., Mignot, A., Guidi, L., Lacour, L., Poteau, A., D'Ortenzio, F., Gentili, B., and Schmechtig, C.: Deep Chlorophyll Maxima in the Global Ocean: Occurrences, Drivers and Characteristics, Global Biogeochemical Cycles, 35, e2020GB006759, https://doi.org/10.1029/2020GB006759, 2021.

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., Emmons, L. K., Fasullo, J., Garcia, R., Gettelman, A., Hannay, C., Holland, M. M., Large, W. G., Lauritzen, P. H., Lawrence, D. M., Lenaerts, J. T. M., Lindsay, K., Lipscomb, W. H., Mills, M. J., Neale, R., Oleson, K. W., Otto-Bliesner, B., Phillips, A. S., Sacks, W., Tilmes, S., van Kampenhout, L., Vertenstein, M., Bertini, A., Dennis, J., Deser, C., Fischer, C., Fox-Kemper, B., Kay, J. E., Kinnison, D., Kushner, P. J., Larson, V. E., Long, M. C., Mickelson, S., Moore, J. K., Nienhouse, E., Polvani, L., Rasch, P. J., and Strand, W. G.: The Community Earth System Model Version 2 (CESM2), Journal of Advances in Modeling Earth Systems, 12, e2019MS001916, https://doi.org/10.1029/2019MS001916, 2020.

Dunne, J. P., Horowitz, L. W., Adcroft, A. J., Ginoux, P., Held, I. M., John, J. G., Krasting, J. P., Malyshev, S., Naik, V., Paulot, F., Shevliakova, E., Stock, C. A., Zadeh, N., Balaji, V., Blanton, C., Dunne, K. A., Dupuis, C., Durachta, J., Dussin, R., Gauthier, P. P. G., Griffies, S. M., Guo, H., Hallberg, R. W., Harrison, M., He, J., Hurlin, W., McHugh, C., Menzel, R., Milly, P. C. D., Nikonov, S., Paynter, D. J., Ploshay, J., Radhakrishnan, A., Rand, K., Reichl, B. G., Robinson, T., Schwarzkopf, D. M., Sentman, L. T., Underwood, S., Vahlenkamp, H., Winton, M., Wittenberg, A. T., Wyman, B., Zeng, Y., and Zhao, M.: The GFDL Earth System Model Version 4.1 (GFDL-ESM 4.1): Overall Coupled Model Description and Simulation Characteristics, Journal of Advances in Modeling Earth Systems, 12, e2019MS002015, https://doi.org/10.1029/2019MS002015, 2020.

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

Hajima, T., Watanabe, M., Yamamoto, A., Tatebe, H., Noguchi, M. A., Abe, M., Ohgaito, R., Ito, A., Yamazaki, D., Okajima, H., Ito, A., Takata, K., Ogochi, K., Watanabe, S., and Kawamiya, M.: Development of the MIROC-ES2L Earth system model and the evaluation of biogeochemical processes and feedbacks, Geoscientific Model Development, 13, 2197-2244, https://doi.org/10.5194/gmd-13-2197-2020, 2020.

Hashioka, T. and Aita, M. N.: Chlorophyll–particle backscattering displacement reveals multiple deep chlorophyll maximum structures in the North Pacific, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2026-3910, 2026a.

Hashioka, T. and Aita, M. N.: Data and code for "CMIP6 biogeochemical models under-represent the deep chlorophyll maximum and diverge in how they express it: a regime-stratified BGC-Argo evaluation in the North Pacific", Zenodo [data set], https://doi.org/10.5281/zenodo.20997630, 2026b.

Kwiatkowski, L., Torres, O., Bopp, L., Aumont, O., Chamberlain, M., Christian, J. R., Dunne, J. P., Gehlen, M., Ilyina, T., John, J. G., Lenton, A., Li, H., Lovenduski, N. S., Orr, J. C., Palmieri, J., Santana-Falcón, Y., Schwinger, J., Séférian, R., Stock, C. A., Tagliabue, A., Takano, Y., Tjiputra, J., Toyama, K., Tsujino, H., Watanabe, M., Yamamoto, A., Yool, A., and Ziehn, T.: Twenty-first century ocean warming, acidification, deoxygenation, and upper-ocean nutrient and primary production decline from CMIP6 model projections, Biogeosciences, 17, 3439-3470, https://doi.org/10.5194/bg-17-3439-2020, 2020.

Lovato, T., Peano, D., Butenschön, M., Materia, S., Iovino, D., Scoccimarro, E., Fogli, P. G., Cherchi, A., Bellucci, A., Gualdi, S., Masina, S., and Navarra, A.: CMIP6 Simulations With the CMCC Earth System Model (CMCC-ESM2), Journal of Advances in Modeling Earth Systems, 14, e2021MS002814, https://doi.org/10.1029/2021MS002814, 2022.

Roesler, C., Uitz, J., Claustre, H., Boss, E., Xing, X., Organelli, E., Briggs, N., Bricaud, A., Schmechtig, C., Poteau, A., D'Ortenzio, F., Ras, J., Drapeau, S., Haëntjens, N., and Barbieux, M.: Recommendations for obtaining unbiased chlorophyll estimates from in situ chlorophyll fluorometers: A global analysis of WET Labs ECO sensors, Limnology and Oceanography: Methods, 15, 572-585, https://doi.org/10.1002/lom3.10185, 2017.

Sauzède, R., Schmechtig, C., Renosh, P. R., Uitz, J., and Claustre, H.: Global Look-Up Table of Physiological Ratios for the Real-Time Adjustment of Chlorophyll-a Fluorescence within the OneArgo Framework, SEANOE [data set], https://doi.org/10.17882/105732, 2026.

https://doi.org/10.5194/egusphere-2026-4562
Preprint. Discussion started: 24 August 2026

Sellar, A. A., Jones, C. G., Mulcahy, J. P., Tang, Y., Yool, A., Wiltshire, A., O'Connor, F. M., Stringer, M., Hill, R., Palmieri, J., Woodward, S., de Mora, L., Kuhlbrodt, T., Rumbold, S. T., Kelley, D. I., Ellis, R., Johnson, C. E., Walton, J., Abraham, N. L., Andrews, M. B., Andrews, T., Archibald, A. T., Berthou, S., Burke, E., Blockley, E., Carslaw, K., Dalvi, M., Edwards, J., Folberth, G. A., Gedney, N., Griffiths, P. T., Harper, A. B., Hendry, M. A., Hewitt, A. J., Johnson, B., Jones, A., Jones, C. D., Keeble, J., Liddicoat, S., Morgenstern, O., Parker, R. J., Predoi, V., Robertson, E., Siahaan, A., Smith, R. S., Swaminathan, R., Woodhouse, M. T., Zeng, G., and Zerroukat, M.: UKESM1: Description and Evaluation of the U.K. Earth System Model, Journal of Advances in Modeling Earth Systems, 11, 4513-4558, https://doi.org/10.1029/2019MS001739, 2019.

Séférian, R., Berthet, S., Yool, A., Palmiéri, J., Bopp, L., Tagliabue, A., Kwiatkowski, L., Aumont, O., Christian, J., Dunne, J., Gehlen, M., Ilyina, T., John, J. G., Li, H., Long, M. C., Luo, J. Y., Nakano, H., Romanou, A., Schwinger, J., Stock, C., Santana-Falcón, Y., Takano, Y., Tjiputra, J., Tsujino, H., Watanabe, M., Wu, T., Wu, F., and Yamamoto, A.: Tracking Improvement in Simulated Marine Biogeochemistry Between CMIP5 and CMIP6, Current Climate Change Reports, 6, 95-119, https://doi.org/10.1007/s40641-020-00160-0, 2020.

Tanner, E. P., Ewin, H., Viljoen, J. J., and Brewin, R. J. W.: Revisiting the relationship between surface and column-integrated chlorophyll-a concentrations in the Biogeochemical-Argo and satellite era, Frontiers in Remote Sensing, 5, 1495958, https://doi.org/10.3389/frsen.2024.1495958, 2024.

Yasunaka, S., Ono, T., Sasaoka, K., and Sato, K.: Global distribution and variability of subsurface chlorophyll a concentrations, Ocean Science, 18, 255-268, https://doi.org/10.5194/os-18-255-2022, 2022.

Yukimoto, S., Kawai, H., Koshiro, T., Oshima, N., Yoshida, K., Urakawa, S., Tsujino, H., Deushi, M., Tanaka, T., Hosaka, M., Yabu, S., Yoshimura, H., Shindo, E., Mizuta, R., Obata, A., Adachi, Y., and Ishii, M.: The Meteorological Research Institute Earth System Model Version 2.0, MRI-ESM2.0: Description and Basic Evaluation of the Physical Component, Journal of the Meteorological Society of Japan. Ser. II, 97, 931-965, https://doi.org/10.2151/jmsj.2019-051, 2019.

Ziehn, T., Chamberlain, M. A., Law, R. M., Lenton, A., Bodman, R. W., Dix, M., Stevens, L., Wang, Y., and Srbinovsky, J.: The Australian Earth System Model: ACCESS-ESM1.5, Journal of Southern Hemisphere Earth Systems Science, 70, 193-214, https://doi.org/10.1071/ES19035, 2020.