A Multi-Variable Class-4 Verification-Uncertainty Budget for Operational Ocean-Forecast Skill with CMEMS In Situ Observations

Sang-Hun Jeong[1], Jae-Il Kwon[1], Ki-young Heo[1]

[1]*Korea Institute of Ocean Science and Technology*

Abstract

Class-4 verification of operational ocean forecasts against observations underpins the OceanPredict IV-TT intercomparison. The resulting skill also depends on documented verification choices, including the quality-control (QC) procedure, sampling-density weighting, reference climatology, scoring rule, and forecast-lead alignment. As a case study for the evaluation protocol of K-Ocean Predict (KOP), a national global-to-coastal prediction system, we quantify their combined effect as a verification-uncertainty budget for two systems (CMEMS, HYCOM ESPC-D) evaluated against the same surface-temperature observations from some 7,400 platforms via the CMEMS In Situ TAC, at a fixed lead over a three-month 2026 record (1.6 million match-ups).

Across five operational QC procedures, the monthly observation-weighted RMSE range reached 0.8–1.3 °C, tens to hundreds of times the inter-system difference, whose confidence intervals included zero in every month-by-lead combination. Accepting provider flags by default produced the largest errors, and the suspect-disposition choice within QARTOD alone drove a monthly spread of 10–20 times that difference. The nominal ranking reversed with month and lead, and the verification-choice spread exceeded platform-block bootstrap uncertainty. Across five further observation types, with provider-rejected records restored as a diagnostic counterfactual, outlier-control sensitivity depended on the observation stream (provider and processing level) rather than the variable, ranging from catastrophic for Argo near-real-time salinity (under 3% of records carrying 97% of the squared error) through dominant for surface temperature to structurally absent for level-3 sea-level anomaly. Within-bin observation spread correlated with match-up error, supporting density control as a representativeness proxy; the budget of an illustrative composite index roughly doubled with weighting alone.

A single skill score did not support ranking the two systems. These results quantify the value of fixing verification choices within intercomparison protocols and suggest that Class-4-type skill scores derived outside them be reported as a range conditioned on a minimum set of documented choices, including the observation source.