WP-MIP White Paper

An Artificial Intelligence, Hybrid and Physically Based Model Intercomparison Project for Weather Prediction

Table of Contents

Version 1.0 (June 2025)

Co-Leads: Linus Magnusson (ECMWF) and Ron McTaggart-Cowan (ECCC)

Background

Advances in artificial intelligence have led to the recent emergence of skillful AI-based and hybrid AI/physically based models for weather prediction applications. Development and evaluation of the performance of these systems has been undertaken by individual institutions, private corporations and academic researchers. These efforts have led to the creation of data aggregation and dissemination platforms; however, a model intercomparison project (MIP) with broad international engagement is needed to facilitate development and evaluation activities at national meteorological centres. As a WMO entity, the Working Group on Numerical Experimentation (WGNE) is well positioned to lead such a global effort.

The WMO Integrated Processing and Prediction System (WIPPS) is a global network of operational centres providing earth-system analyses and predictions to all WMO Members and the wider community. Designated WIPPS centres agree to provide defined sets of mandatory and recommended products to all Members. Developments in AI may have significant implications for operational practices and the evolution of the WIPPS. Artificial intelligence-based systems are significantly cheaper to run than traditional NWP models and bring substantial opportunities as well as potential risks. A priority is to provide guidance to users on the use of AI-based forecasts. An intercomparison of data-driven models and comparison of strengths and weaknesses compared to traditional NWP models will provide essential guidance to WMO Members.

The need for an AI-inclusive MIP was identified by WIPPS as a pilot project proposed to WGNE at its 39th Annual Meeting (Fall 2024). Discussions at the WGNE meeting revealed broad support for this initiative within the numerical modelling community and laid out the primary objectives for the project.

Development and fine-tuning of AI-based models requires access to significant technical and computing resources, stressing the capacities of smaller operational centres. Despite these challenges, all members of WGNE agree that broad engagement in an AI-focused MIP is essential, particularly given questions about the performance of AI-based models in regions with sparse observations. A key project outcome will therefore be guidance for assessments of prediction quality, an objective that will be achievable through the active involvement of the Joint Working Group for Forecast Verification Research (JWGFVR).

For ease of reference, we hereafter follow the terminology of Radford et al. (2024) in which AI-based models are referred to as AIWP (artificial intelligence weather prediction) systems to distinguish them from physically based NWP models. Both are specific instances of weather prediction (WP) systems, an umbrella designation that gives rise to the proposed intercomparison project's name: WP-MIP.

Context and Coordination

The goals of the proposed MIP have been established to complement existing efforts in AIWP model development and evaluation rather than duplicating them. An overview of the objectives of relevant projects is shown in Fig. 1, with the clearly distinguishing features of the two proposed streams of WP-MIP (dark and light blue) being their focus on the broad engagement of national weather services and a scope that includes the full spectrum of guidance including NWP, hybrid and AIWP systems.

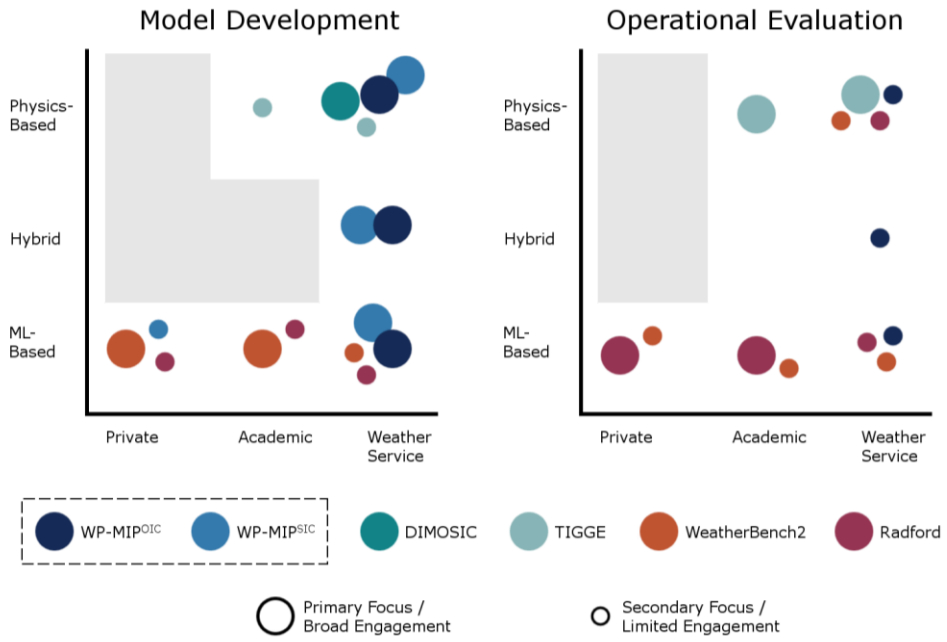

Figure 1. *Overview of objectives and engagement for current and recent relevant projects as identified in the legend below the panels. Model development (left) and operational evaluation (right) objectives are shown independently for different classes of weather prediction models (ordinate) developed and evaluated by different sources (abscissa). The size of the marker reflects both priority and breadth of engagement as indicated in the secondary legend. The greyed-out regions on the panels represent system-contributor combinations that are not expected to be dominant in the near future. Project names are defined individually in the text.*

The existing WeatherBench2 project (Rasp et al. 2024) focuses on providing benchmark datasets for pure AIWP model development and headline evaluation tools for leaderboard-type assessments (orange in Fig. 1). Its current iteration centres on 2020 forecasts; however, an update to 2022 is planned. This platform serves as a valuable resource for AIWP model developers, providing easy access to benchmarking datasets and evaluation tools. Led primarily by Google Research, WeatherBench2 has a close connection to ECMWF for benchmark and evaluation datasets. Although a protocol for inclusion of data from other sources has been developed, engagement appears to have been primarily with private firms.

The limited involvement of other national meteorological centres may be related to difficulties in interpreting inferences initialized from non-ECMWF data and the restrictions on fine-tuning implied by an older evaluation period.

The AIWP accelerator project described by Radford et al. (2024) takes a more operational approach by prioritizing real-time data and visualization products (red in Fig. 1). One of the important goals of this project is to build forecaster experience with AIWP and confidence in the inferences that it provides. As a NOAA-supported effort, Radford et al. (2024) focus on GFS-initialized pure AIWP models that have been contributed to the project primarily from American and private sources. Although the project also includes a three–year archive for inferences from some of these models – limited in some cases by fine-tuning periods – the key objectives of the project focus on evaluation of mature AIWP models to pave the way for eventual NOAA operationalization.

The proposed WP-MIP project complements these efforts through its emphasis on engagement across national meteorological centres on the full spectrum of weather prediction systems (light and dark blue in Fig. 1). By doing so, it leverages the investments made over decades of model development and evaluation at these weather services in the spirit of The International Grand Global Ensemble (TIGGE; cyan in Fig. 1) but with significantly different objectives. Participants will be encouraged to evaluate all contributions using their standard tools and a broad range of observations, and to develop novel evaluation techniques to assess the quality of guidance produced by the systems. Active operational community involvement will also facilitate access to a variety of analyses for both initialization and evaluation. Using a recent period (2024) will encourage contributions from fine-tuned AIWP and hybrid systems and facilitate all forms of model development by allowing for quantification of relative strengths and weaknesses throughout the development cycle.

In order to encourage participation and to leverage the community's investment in the recent DIfferent MOdels Same Initial Conditions (DIMOSIC; green in Fig. 1) project (Magnusson et al. 2022) – the second phase of which was expected to begin early in 2024 – two coordinated streams will be implemented within WP-MIP (Fig. 2). These streams are distinguished by the model initializations: the original DIMOSIC strategy (common ECMWF initializations) will be employed for the Same Initial Conditions (SIC) stream (light blue in Figs. 1 and 2), while each centre's individual analyses will be used to initialize models contributed to the Own Initial Conditions (OIC) stream (dark blue in Figs. 1 and 2). Employing common data handling and headline assessment strategies will accelerate the proposed project by leveraging both the tools and experience developed during DIMOSIC. This close coordination will also facilitate initialization sensitivity studies between WP-MIP streams.

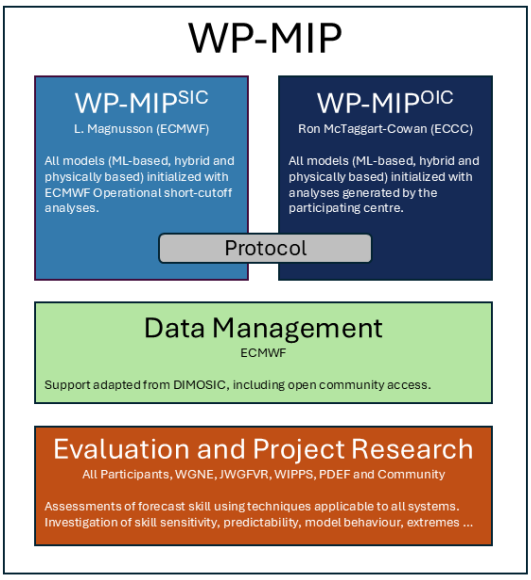

***Figure 2**. Design of proposed WP-MIP project.*

In addition to the two "core" streams of the project, subprojects can be introduced to augment the WP-MIP archive for specific needs. These subprojects (denoted "SP#") can be proposed and organized by interested researchers, with the data being added to the main project archive. Figure 3 summarizes subproject integration into WP-MIP. Groups contributing simulations to the core streams project can choose which of these subprojects they wish to contribute to. The subprojects are described in Appendix A at the end of this document.

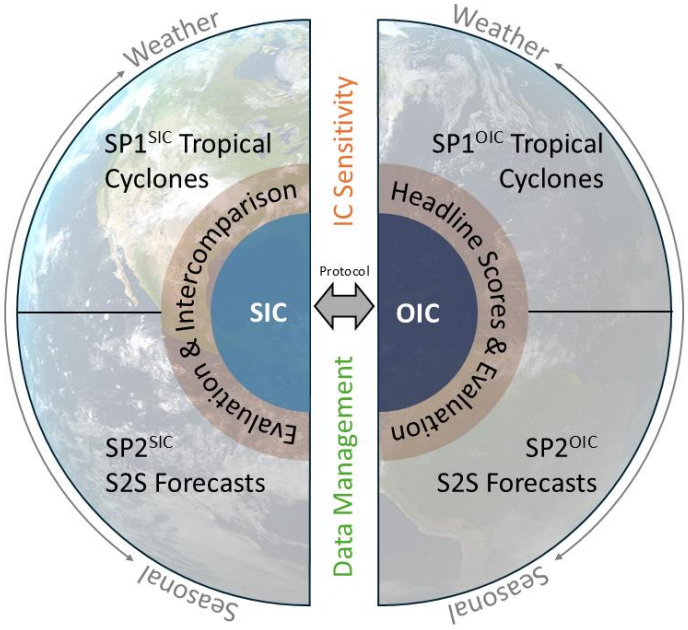

Figure 3. WP-MIP organization and subproject integration. Colour coding of project components follows that of Fig. 2.

Objectives

The WP-MIP project aims to gather medium-range predictions from the full range of modelling systems (NWP, hybrid and AIWP) following a common experimental protocol. These data will be made publicly available for research and model evaluation projects.

The current generation of AIWP models may suffer from a number of important deficiencies that could have negative impacts on their uptake and practical utility. An important outcome of WP-MIP will be to provide the datasets required to investigate the validity of these concerns and their impact on predictions:

- smoothing of predicted fields;
- stability of AIWP models;
- physical consistency, dynamical balance and conservation properties in AIWP system;
- effective resolution of pure AI models, including comparison to hybrid systems;
- quality of precipitation forecasts;
- ability of AIWP systems to predict extreme or out-of-sample events; and,
- utility of tropical cyclone track and intensity guidance.

The WP-MIP objective to study physical consistency and dynamical balance will be used to inform ongoing discussions within the WMO Task Team on Reviewing NWP Standardized Verification (TT-NWPSV) around effective verification methods for NWP and AIWP systems. With the support of the JWGFVR, WP-MIP results will be used to determine the ability of different verification techniques to identify strengths and weaknesses in the project database. One outcome of this effort may be a checklist of physical consistency tests that could be employed during both model development and operational evaluation.

The sensitivity of AIWP models to initial conditions remains another outstanding scientific question that is highly relevant to the fine-tuning activities currently underway at many operational centres. The DIMOSIC project made it clear that NWP models benefitted from high-quality initial conditions even if they were generated by an “alien” data assimilation system (Magnusson et al. 2022). The extent to which this conclusion will hold for AIWP models is unknown.

The two streams of WP-MIP will provide complementary information that will be useful in different applications. Developers of NWP models will likely focus on the SIC stream as a way to reduce uncertainty in the initial conditions of the system, thereby making the model itself the primary source of forecast error. Researchers focusing on AIWP will likely make use of both streams because both the operationalization of locally trained and fine-tuned systems (OIC) and

the assessment of initial condition sensitivities (cross-stream comparison) remain outstanding questions.

Identifying the precise origin of the impressive predictive skill of AIWP models represents a significant research challenge. Participation of the Predictability, Dynamics and Ensemble Forecasting (PDEF) working group in WP-MIP will ensure that studies aimed at developing an understanding of the source of AIWP skill will be an important component of the project. Such investigations will help to build confidence in the guidance generated by these systems in addition to addressing fundamental questions about atmospheric predictability and the effectiveness of AIWP ensembles in quantifying uncertainty.

Broad participation in WP-MIP is a requirement for its success. This means that the experimental protocol should be flexible enough to encourage contributions while being sufficiently controlled to allow for meaningful comparisons on these and other emerging topics. The initial phase of WP-MIP will focus on global modelling, potentially with future iterations taking on specific regional foci.

Methods for evaluation of AI and hybrid models are currently under development. Active involvement in WP-MIP by the JWGFVR is essential to ensure that the latest verification techniques are brought to bear on the problem, and to make the project dataset a testbed for further verification research.

Core Protocol

The WP-MIP protocol is inspired by that of the DIMOSIC project, with WP-MIP^{SIC} representing its direct successor – with extension to AIWP and hybrid systems – and WP-MIP^{OIC} an extension to centre-specific initializations. This design will encourage participants to contribute to both streams of this intercomparison effort. Crossover between streams will facilitate studies of initial condition sensitivity and predictability more generally.

Project Period

Model development groups typically use recent periods to run hindcasts for model evaluation. This familiarity makes this protocol very recognizable for developers, who have well-established tools and data streams for model integration and assessment. An additional requirement for this protocol is that the hindcast periods be very recent to avoid problems with AIWP systems that use rapid fine-tuning cycles that may overlap with older hindcasting periods and to allow the models to be initialized with analyses from the most up-to-date data assimilation system.

The hindcast period(s) should be long enough to sample broadly from synoptic-scale variability. As noted by both Rasp et al. (2024) and Magnusson et al. (2022) projects, a 1-year period – although insufficient to sample extremes and low-frequency variability robustly – appears to be sufficient for a stable evaluation. **1 January 2024 through 31 December 2024** would provide

results from the most recent analysis systems and minimize the potential for training contamination. Initializations should be made at 0000 UTC at 3-day intervals to minimize the computational cost of the protocol (particularly for hybrid and NWP models) and maximize independence at the synoptic scales.

For the WP-MIP^{SIC} stream, models should be initialized with full-resolution hybrid-coordinate operational ECMWF analyses if possible. For hybrid or AIWP systems that cannot ingest these data, pressure-coordinate analyses may be degraded to the appropriate horizontal grid spacing. Full-resolution, hybrid- and pressure-coordinate analyses will be made available to project participants, who will be responsible for their own aggregation (if necessary) for model initialization.

For the WP-MIP^{OIC} stream, models should be initialized with the operational analyses employed at the participating centre. This means that the data from the NWP model will likely be generated from the current operational system at each centre. Using locally generated initialization will make the WP-MIP^{OIC} dataset directly relevant for evaluation of operationally deployable guidance (Fig. 1).

Use of a common, well-defined period will facilitate the addition of new models as they emerge and will allow for the development of baseline forecast and observational datasets. Comparison of results between groups will also be more direct when a prescribed period is used. Participating models must not include the hindcast period in their AIWP and hybrid-model training to ensure a fair assessment. It must therefore be anticipated the future iterations of WP-MIP will need to employ more recent project periods.

The primary disadvantage of this hindcast protocol is that operational forecaster engagement will be minimal. Some centres may have the capacity to dedicate forecaster resources to WP-MIP assessment as a special project, but most will not. Similarly, WIPPS participation in this protocol will likely be minimal because of its research focus. To mitigate this problem, ECMWF has agreed to provide an extreme event catalog for the project period, which could be used to encourage forecaster engagement for specific high-impact case studies. The Extreme Weather Bench project proposed by Brightband may also serve as a valuable data source for this purpose.

Model Deliverables

A common set of outputs is prescribed in this protocol to facilitate archiving, evaluation and diagnosis; however, not all models will be able to generate the full set of requested fields. For example, most current AIWP models generate outputs on a relatively limited set of pressure levels for a small number of state variables. Contributors should provide the most complete dataset possible for each of their modelling systems.

Participating groups should submit results on 17 mandatory levels (1000 hPa, 925 hPa, 850 hPa, 700 hPa, 500 hPa, 400 hPa, 300 hPa, 250 hPa, 200 hPa, 150 hPa, 100 hPa, 70 hPa, 50 hPa , 30 hPa, 20 hPa, 10 hPa, 1 hPa), or the largest subset of these levels possible:

- Dry air temperature (K)
- Specific humidity (kg kg^{-1})
- Zonal and meridional wind components (m s^{-1})
- Geopotential height (m)

Additionally, several 2D fields should be provided if they are produced by the participating model, with accumulated fields accumulating since the beginning of the run (no resets on output):

- Sea level pressure (hPa)
- Surface pressure (hPa)
- Screen-level (2 m) temperature (K)
- Screen-level (2 m) specific humidity (kg kg^{-1})
- Screen-level dewpoint (K)
- Anemometer-level (10 m) zonal and meridional wind components (m s^{-1})
- Precipitation accumulation since the beginning of the integration (kg m^{-2})
- Sea surface temperature (K)
- Sea ice cover (fractional)
- Total (2D) cloud cover (%)
- Low-level cloud cover (%)
- Mid-level cloud cover (%)
- High-level cloud cover (%)
- Accumulated net longwave radiation flux at the top of the atmosphere (W m^{-2} · s)
- Accumulated net solar radiation flux at the top of the atmosphere (W m^{-2} · s)
- Accumulated downwards longwave radiation flux at the surface (W m^{-2} · s)
- Accumulated net longwave radiation flux at the surface (W m^{-2} · s)
- Accumulated downwards solar radiation flux at the surface (W m^{-2} · s)
- Accumulated net solar radiation flux at the surface (W m^{-2} · s)
- Accumulated Surface turbulent sensible heat flux (W m^{-2} · s)
- Accumulated Surface turbulent latent heat flux (W m^{-2} · s)

Finally, invariant/fixed fields should be provided in the 0 h forecast of each integration:

- Land-sea mask (fixed; fraction)
- Orography (fixed; m)

The focus of WP-MIP on medium-range predictions on the global domain suggests that outputs should be provided at 6-hourly intervals for 10-day forecasts. Although this will only minimally sample the diurnal cycle, time-stepping limitations of current AIWP models means that this output frequency represents a reasonable request.

The initial phase WP-MIP will focus on global deterministic predictions, with most outputs exchanged on a 0.25° lat/lon grid. Interpolation or aggregation should be performed by the contributor prior to uploading data. Future iterations of WP-MIP may be extended to ensemble predictions, which will be of particular interest for assessments of extremes. A limited subset of integrations may also be exchanged on a higher resolution grid to facilitate assessments of smoothing (energy spectra, activity metrics, etc) and tropical cyclone intensity.

In addition to the forecasts, centres contributing to the WP-MIP^{OIC} stream should submit their highest quality (long-cutoff) analyses for use in evaluation activities. These analyses should also be aggregated onto the 0.25° lat/lon exchange grid and provided at 6-hourly intervals for the full project verification period (1 January 2024 to 8 January 2025).

Data Formatting

The ECMWF already maintains archives of both experimental and operational forecasts as part of the TIGGE and Subseasonal to Seasonal (S2S) projects, in addition to providing access to inferences from AIWP models run within the centre. The WP-MIP exchange will take advantage of the protocols and infrastructure already built for TIGGE to facilitate contributions. This includes making the data available on the MARS system for exchange and community access.

The data formatting requirements for WP-MIP are modeled directly on the TIGGE guidelines for data encoding and exchange with the following exceptions:

- all data should be provided on a 0.25° lat/lon grid with points at the poles and no wrapping at 0°E, so 1440x721 grid points;
- the production status of delivered data should be 2 (research products) following GRIB2 Table 1.3;
- the data type value should be 1 (forecast product) following GRIB2 Table 1.4;

The file naming format should be:

wpmip_SSS_CCCC_YYYYMMDDHH_MM_II_LL.grib2

- SSS: oic/sic/tce/s2s subproject (i.e. OIC/SIC/tc evaluation/S2S as defined for subprojects)
- CCCC: centre acronym (contributors without an acronym should use their own four-letter identifiers after confirming that they do not conflict)
- YYYYMMDDHH: date * time stamp (e.g. 2024100100 for 0000 UTC run on 1 October 2024)
- MM: pm/ai/hy model type (i.e. physical/ai/hybrid)
- II: model iteration number starting from 00 (allows multiple versions of a model)
- LL: pl/sl/pt/pv (level type i.e. pressure/surface/pv level/pt)

Grib2 Descriptions for Pressure-Level Fields[1]

Parameter	Category (Table 4.1)	Parameter (Table 4.2)	Level (Table 4.5)	Units
Temperature	0	0	100	K
Specific Humidity	1	0	100	kg kg^{-1}
U-Component Wind	2	2	100	m s^{-1}
V-Component Wind	2	3	100	m s^{-1}
Geopotential Height	3	5	100	m

Grib2 Descriptions for Single-Level Fields[1]

Parameter	Discipline (Table 0.0)	Template (Table 4.0)	Category (Table 4.1)	Parameter (Table 4.2)	Level (Table 4.5)	Processing (Table 4.10)	Units
Sea Level Pressure	0	0	3	1	101	0	Pa
Surface Pressure	0	0	3	0	1	0	Pa
Screen-Level Temperature	0	0	0	0	103	0	K
Screen-Level Specific Humidity	0	0	1	0	103	0	kg kg^{-1}
Screen-Level Dewpoint	0	0	0	6	103	0	K
Anemometer-Level U-Component Wind	0	0	2	2	103	0	m s^{-1}
Anemometer-Level V-Component Wind	0	0	2	3	103	0	m s^{-1}

[1] Discipline 0 (Meteorological Products) as per Grib2 Table 0.0.

Sea Surface Temperature	10	0	3	0	1	0	K
Sea Ice Cover	10	0	2	0	1	0	fraction
Accumulated Precipitation	0	8	1	52	1	1	Kg m^{-2}
Total (2D) cloud cover	0	0	6	1	10	0	%
Low cloud cover	0	0	6	3	10	0	%
Mid cloud cover	0	0	6	4	10	0	%
High cloud cover	0	0	6	5	10	0	%
Accumulated Net Longwave Radiation (TOA)	0	8	5	5	8	1	W m^{-2} · s
Accumulated Net Shortwave Radiation $(TOA)^2$	0	8	4	9	8	1	W m^{-2} · s
Accumulated Downwards Longwave Radiation (Surface)	0	8	5	3	1	1	W m^{-2} · s
Accumulated Net Longwave Radiation (Surface)	0	8	5	5	1	1	W m^{-2} · s
Accumulated Downwards Shortwave Radiation (Surface)	0	8	4	7	1	1	W m^{-2} · s
Accumulated Net Shortwave Radiation (Surface)	0	8	4	9	1	1	W m^{-2} · s
Accumulated Surface Turbulent	0	8	0	11	1	1	W m^{-2} · s

Sensible Heat Flux[2]							
Accumulated Surface Turbulent Latent Heat Flux[2]	0	8	0	10	1	1	$W\ m^{-2} \cdot s$

Grib2 Descriptions for Fixed Fields

Parameter	Discipline (Table 0.0)	Category (Table 4.1)	Parameter (Table 4.2)	Level (Table 4.5)	Units
Land-Sea Mask	2	0	0	1	None
Orography	2	0	7	1	m

Data Exchange

Instructions for uploading WP-MIP project data to MARS will be distributed to the contributors as needed.

Evaluation and Project Research

All participants will have access to the full dataset for their own diagnostic and evaluation studies. Although some of these will be aimed at addressing the WP-MIP objectives, all others will also be welcome additions to the project.

Standard headline scores will be computed against a set of operational analyses/reanalyses, and the global radiosonde and synoptic station networks (the latter for upper-air and near-surface variables, respectively). Standard verification against point observations (radiosondes and synop stations) will follow the currently reviewed WIPPS verification standards for Deterministic Prediction Systems (WMO manual 485, Appendix 2.2.34), with sole exception of interpolating the forecast values to the station location from the common WP-MIP 0.25° lat/lon grid. However, adjustment to this directive might follow after a spectral analysis, if participants wish to render all predictions to the same effective resolution.

Assessment of WIPPS-identified upper-air and near-surface variables (the latters being possibly more user-relevant) is essential for providing guidance to National Weather Centers on the capabilities of the different systems (e.g. if AI-systems have weaknesses in representing

[2] Unless otherwise specified, fluxes are positive downwards.

precipitation, a key variable for the public, this needs to be known). A desired outcome from this exercise is a document with recommendations for operational verification of WP, encompassing AI, hybrid and physically-based systems.

Verification against own analysis is affected by the dependence between the analysis and prediction system. While this is well known in the NWP community (Parks, 2008), the effect of the analysis incestuousness has not been yet rigorously explored for AI models (most trained on ECMWF reanalyses). Verification against own and independent analysis is hence proposed, where the independent analysis should be a multi-center ensemble or similar product that allows for quantification of analysis uncertainty during evaluation.

The JWGFVR will play an important role in identifying and implementing novel evaluation techniques that can be applied to the WP-MIP forecast database. These could include:

- Scale-separation techniques (e.g.Casati et al 2023; Buschow and Friederichs, 2020, 2021). Traditional methods, such point-by-point means squared error, tend to over-penalize detailed high-resolution forecasts, compared to their smoother coarse-resolution counterparts, because of their higher variability and double penalties due to small scale displacements in the higher-resolution gridded products. The scale-separation methods will be used to filter predictions, facilitating fair comparisons between smoother AI-models and more detailed NWP models (e.g. BenBouallegue et al 2023; Husain et al, 2024). Moreover, these methods permit the identification of performance strengths or prediction weaknesses for physically meaningful weather scales separately (e.g. planetary, synoptic and meso-scales, as in Jung and Leutbecher, 2008), which is more informative than summary statistics.
- Feature-based and optimal transport / field morphing techniques (e.g. Keil and Craig 2007, 2009; Marzban and Sandgathe 2010; Skok 2023) address directly the double penalty issue by scoring phase and amplitude error – and displacement and intensity errors – separately (e.g. for precipitation, as in Davis et al 2006). These techniques will likely play an important role in WP-MIP evaluation because they could also serve to improve spectral nudging in hybrid systems.
- Process-oriented diagnostics and physical coherence assessment are required, since AI-models (as opposed to physically based NWP) do not intrinsically guarantee physical coherence between variables. Public forecasts delivered daily by National Weather Centers cannot display inconsistencies (e.g. snow with positive temperatures); hence National Weather Centers need to be informed on the AI-models physical consistency, prior fully investing in AI for their forecast products. As a corollary, this assessment could also help identify processes mis-representations in physically based NWP systems.
- Rigorous evaluation of extreme events is necessarily complicated by their relative infrequency, a problem that is made more acute by the 1-year period of the WP-MIP forecast dataset. Evaluation techniques will be developed to address this limitation, which is likely to persist for AIWP models because of the need to maximize training and fine-tuning dataset sizes. As an initial step, case-study assessments will be based on events identified in the ECMWF extreme event catalog and by the Extreme Weather Bench project (McGovern et al. 2025). Methods that go beyond a simple comparison of

the amplitude of the events in forecast and observation will be encouraged as it is expected that smoother fields would capture less accurately the amplitude of extreme events.

These methods will likely be explored in tailored research projects with possible collaborations with universities.

Resources

The WP-MIP project is unfunded. All resources are provided as in-kind support from participating institutions. The initial list of contributors is based on WGNE membership and preliminary WP-MIP discussions. Additional contributions are more than welcome from operational centres, academic institutions and the private sector. This table should be extended to include participants as needed. It is understood that contributions from non-operational participants will primarily be focused on the WP-MIP^{SIC} stream.

The list provided here includes both forecast contributors and diagnostic teams. Please see the WP-MIP Contribution Dashboard for a real-time accounting of forecast contributions.

Centre	Contact	Data Contribution	Support & Evaluation
BoM	Debbie Hudson	Operational ACCESS-G (TBC)	
		AIFS initialised with ACCESS-G	
CEMC		None	
CPTEC	Caio Coelho	Global operational sub-seasonal model for SP2	
CSIR	Mohau Mateyisi	Global model	Tropical cyclone evaluation and sub-seasonal prediction evaluation.
DWD	Martin Köhler	Operational ICON, later AICON our AI model	Focus on SYNOP and surface/TOA fluxes including precip and energy and water budgets.
ECCC	Ron McTaggart-Cowan	Operational GDPS	Spectral analyses for smoothing assessment
		Spectral nudging to AI inference	

		PARADIS inferences	
ECMWF	Linus Magnusson and Inna Polichtchouk	Operational IFS	Host WP-MIP data archive; headline scores (shared with DIMOSIC2); extreme event catalog for project period (L. Magnusson)
		Spectral nudging to AIFS	
		AIFS inferences	
GFDL	Jan-Huey Chen	SHiELD	Tropical cyclone evaluation
IITM	Ankur Srivastava	None	Monsoon depressions, extreme rainfall, heat
INPE		None	
JMA	Masashi Ujiie	Global model	WGNE tropical cyclone evaluation
JWGFVR	Barbara Casati; Zied Ben Bouallegue	None	Develop methodology to assess physical coherence; investigate potentials of scale-separation and spatial methods.
KIAPS	Eun-Hee Lee; Kyoungmi Cho	Operational KIM	
Met Office	Stephen Haddad	UM Forecasts	Extratropical cyclone tracking and evaluation (D. Ackerley?)
Met Norway	Cristian Lussana	Global AIWP for LAM LBC	Global wavelet-based analysis for scale-separated evaluation
Meteo-France		None	
NCAR		None	
NOAA	Fanglin Yang	Operational GFS	MJO, Monsoon Index
		Experimental ML-GFS	
NRL	Alex Reinecke	Global model	

PDEF	Laure Raynaud	None	Recommendation on methodology, metrics and use cases
RAS	Mikhail Tolstykh	Global 10-km model	
SAWS	Nico Kroese	Global model	
WGSIP	Debbie Hudson	None	Evaluation of results related to the sub-seasonal timescale processes (e.g., MJO, SAM, monsoon…)
WIIPS		None	

Data and Software Policy

Participants should adhere as much as possible to FAIR principles. All data generated as part of the WP-MIP project should be made readily and freely available within the constraints of software licensing agreements. Similarly, all software used in WP-MIP and developed for the project should be shared to the greatest extent possible.

The ECMWF has undertaken to host WP-MIP data on MARS. The dataset can be accessed by all users with a MARS account. Specific information about data access will be provided once preliminary WP-MIP datasets become available.

Timeline

There is no strict timeline for submission of results by participants. However, the use of a fixed hindcast period for WP-MIP means that fine-tuning efforts will progressively move towards the project period. Additional iterations of WP-MIP with updated periods and protocols may be considered in the future if needed.

Nov-Dec 2024: Consultation on WP-MIP White Paper (WGNE, JWGFVR, WIPPS, PDEF, operational centres, academia) and design of WP-MIP protocol (participants).

Feb 2025: Finalize protocol proposal draft and develop technical documentation for the project (participants).

Mar 2025: Virtual meeting of all interested parties to finalize the protocol (1400-1530 UTC 25 March).

2 Apr 2025: “Draft” status of Whitepaper removed.

Apr 2025: Participating centres should develop associated software and processes.

May 2025: Initializing data for SIC stream posted.

July 2025: Finalization of transfer and archiving protocol.

July 2025: Begin production and data transfer for OIC stream.

Sept 2025: Target completion for OIC stream data delivery.

Sept 2025: Target completion for SIC stream data delivery.

Appendix A: WP-MIP Subprojects

Subprojects represent extensions to the WP-MIP protocol described in the main body of this document. Participation in these subprojects is optional, but offers benefits to contributors in the form of additional domain-specific assessments of model behaviour. Potential contributors should contact the relevant subproject lead to inform them of the intention to provide data that follows the subproject protocol. This will ensure that contributions are not overlooked during subproject evaluation and follow-up research.

Subproject SP1 - Tropical Cyclones

Lead: Masashi Ujiie

The goal of SP1 is to complete systematic JMA tropical cyclone evaluation (Yamaguchi et al. 2017). In order to arrive at robust conclusions, this study requires more data than provided for the core streams of WP-MIP.

A second objective of SP1 is to compare results of tropical cyclone tracking between the JMA tracker and the GFDL tracker (Chen et al. 2023). The GFDL tracker is sensitive to the 3D structure of the storm and requires additional data as described in item 3 below. After establishing a tracker comparison baseline using NWP models, sensitivities in hybrid and AIWP predictions will provide additional information about storm structure and how well tropical cyclones in these models align with expected archetypes.

The reduced list of requested outputs (items 2 and 3 below) is intended to ease the data processing and archiving burden for contributing groups. However, sufficient space is available in the MARS archive to accept full core-data contributions for SP1, so some contributors may find it easier simply to provide complete datasets for each date rather than reducing output lists as described here.

Protocol Modification:

1. Runs should be initialized every day rather than every 3 days (366 total runs).
2. Data for the runs that fill in the dates from the standard protocol can be limited to sea level pressure and 10 m wind components (grib2 encoded as for the core protocol).
3. [GFDL tracker inputs] Data for the the runs that fill in the dates from the standard protocol (as for item 2; encoded as for the core protocol):
    a. U- and V-component winds at 850 hPa and 500 hPa
    b. Geopotential height at 850 hPa, 500 hPa and 300 hPa

Subproject SP2 - Sub-Seasonal Forecasts

Lead: WGSIP. Contact: Debbie Hudson

Summary:

The goal of SP2 is to begin assessments of AIWP, physically-based (dynamical) and hybrid models at extended forecast ranges.

Although a complete evaluation of sub-seasonal forecasts requires reforecasts and multi-year periods of ensemble forecasts, this subproject will serve as a preliminary investigation of S2S-relevant features and processes (e.g. MJO, monsoon, SAM). It aims to provide an initial assessment of the performance of forecasts out to week-2, focussing on aspects such as the evolution of biases, physical consistency, and case studies of significant events.

Protocol Modification:

1. Runs should be extended to 15 days (rather than the 10 days required by the core protocol).
2. Analyses should be provided for the 5 additional evaluation days at the end of the period (e.g. to evaluate the last forecast initialized on 29 December 2024, include analyses for 0000 UTC 8 January to 0000 UTC 13 January 2025 as an extension to the core protocol).

Additional Background:

The design of the WP-MIP is not ideal for S2S evaluation. It involves deterministic models, an extremely short period of forecasts and there are no hindcasts/reforecasts. These are all major limitations for assessing forecasts on sub-seasonal timescales.

However, this MIP will be the first resource of ML model forecasts where the models have been run according to a common protocol. It therefore represents an opportunity to get an initial appreciation of the strengths and weaknesses of a range of ML models for capturing S2S-relevant features and processes. This is just a start for S2S evaluation. There is no doubt that more relevant S2S MIPS will emerge very soon as the development of ML models targeting the sub-seasonal timescale progresses. In addition, it is likely that there will be future iterations of this WP-MIP, for example expanded to include ensembles.

Subproject SP2 should be viewed as being not so much about forecast skill (or which model is better than the other), but as a preliminary opportunity to assess how a variety of ML models capture some S2S processes. Using it as a resource for case studies may be the best approach and it can be used to potentially augment existing/planned studies. For example, if a science study is looking at a particular case (e.g. an MJO event in 2024, or the Jan 2024 Arctic SSW) with dynamical models – then it may also be useful to see how the ML models capture aspects of the event (for example, one might look at a set or ensemble of ML models compared with a set or ensemble of physical models for the event). We may also, for example, be able to aggregate processes (e.g. MJOs) across all forecasts and evaluate how the biases evolve in the ML models and how that differs from the dynamical models. Much of this analysis will also be possible using data from the Core Contributions of the MIP, but the extension to 15-days will provide some additional lead time information.

References

Ben Bouallègue, Z., and Coauthors, 2024: The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context. *Bull. Amer. Meteor. Soc.*, **105**, E864–E883, https://doi.org/10.1175/BAMS-D-23-0162.1.

Buschow S, Friederichs P. SAD: Verifying the scale, anisotropy and direction of precipitation forecasts. *QJR Meteorol Soc*. 2021; 147: 1150–1169. https://doi.org/10.1002/qj.3964

Buschow, S. and Friederichs, P. (2020) Using wavelets to verify the scale structure of precipitation forecasts. *Advances in Statistical Climatology, Meteorology and Oceanography*, 6(1), 13–30. https://doi.org/10.5194/ascmo-6-13-2020

Casati, B., Lussana, C., & Crespi, A. (2023). Scale-separation diagnostics and the Symmetric Bounded Efficiency for the inter-comparison of precipitation reanalyses. *International Journal of Climatology*, 43(5), 2287–2304. https://doi.org/10.1002/joc.7975

Chen, J.-H., L. Zhou, L. Magnusson, R. McTaggart-Cowan, and M. Köhler, 2023: Tropical cyclone forecasts in the DIMOSIC project—Medium range forecast models with common initial conditions. *Earth Space Sci.*, **10**, e2023EA002821, https://doi.org/10.1029/2023EA002821.

Davis, C., Brown, B. and Bullock, R. (2006) Object-based verification of precipitation forecasts. Part I: methodology and application to mesoscale rain areas. *Monthly Weather Review*, 134(7), 1772–1784. https://doi.org/10.1175/MWR3145.1

Keil, C., and G. C. Craig, 2007: A Displacement-Based Error Measure Applied in a Regional Ensemble Forecasting System. *Mon. Wea. Rev.*, **135**, 3248–3259, https://doi.org/10.1175/MWR3457.1.

Keil, C., and G. C. Craig, 2009: A Displacement and Amplitude Score Employing an Optical Flow Technique. *Wea. Forecasting*, **24**, 1297–1308, https://doi.org/10.1175/2009WAF2222247.1.

Jung, T. and Leutbecher, M. (2008), Scale-dependent verification of ensemble forecasts. Q.J.R. Meteorol. Soc., 134: 973-984. https://doi.org/10.1002/qj.255

Husain S.Z., L. Separovic, J.-F. Caron, R. Aider, M. Buehner, S. Chamberland, E. Lapalme, R. McTaggart-Cowan, C. Subich, P. Vaillancourt, J. Yang, and A. Zadra (2024): Leveraging data-driven weather models for improving numerical weather prediction skill through large-scale spectral nudging, July 2024. arXiv:2407.06100 [physics]. URL http://arxiv.org/abs/2407.06100

Magnusson, L., and Coauthors, 2022: Skill of Medium-Range Forecast Models Using the Same Initial Conditions. *Bull. Amer. Meteor. Soc.*, **103**, E2050–E2068, https://doi.org/10.1175/BAMS-D-21-0234.1.

McGovern, A., D. Rothenberg, N. Loveday, C. K. Potvin, E. Gilleland and L. H. Kim, 2025: Extreme Weather Bench. *105th Annual Meeting of the American Meteorological Society*, New Orleans, USA, Amer. Meteor. Soc., https://ams.confex.com/ams/105ANNUAL/meetingapp.cgi/Paper/451220.

Radford, J. T., I. Ebert-Uphoff, J. Q. Stewart, K. D. Musgrave, R. DeMaria, N. Tourville, and K. Hilburn, 2024: Accelerating Community-Wide Evaluation of AI Models for Global Weather Prediction by Facilitating Access to Model Output. *Bull. Amer. Meteor. Soc.*, https://doi.org/10.1175/BAMS-D-24-0057.1, in press.

Rasp, S. and Coauthors, 2024: WeatherBench 2: A benchmark for the next generation of data-driven global weather models. Preprint available on arxiv at https://arxiv.org/pdf/2308.15560 .

Marzban, C., and S. Sandgathe, 2010: Optical Flow for Verification. *Wea. Forecasting*, **25**, 1479–1494, https://doi.org/10.1175/2010WAF2222351.1.

Skok, G.: Precipitation attribution distance, Atmos. Res., 295, 106998, https://doi.org/10.1016/j.atmosres.2023.106998, 2023. a

Yamaguchi, M., J. Ishida, H. Sato, and M. Nakagawa, 2017: WGNE Intercomparison of Tropical Cyclone Forecasts by Operational NWP Models: A Quarter Century and Beyond. *Bull. Amer. Meteor. Soc.*, **98**, 2337–2349, https://doi.org/10.1175/BAMS-D-16-0133.1.