TECHNICAL MEMORANDUM

829

The development and evaluation process followed at ECMWF to upgrade the Integrated Forecasting System (IFS)

R Buizza, M Alonso Balmaseda, A Brown, S English, R Forbes, A Geer, T Haiden, M Leutbecher, L MAgnusson, M Rodwell, M Sleigh, T Stockdale, F Vitart and N Wedi

Research Department

This document has been presented at the 47th Scientific Advisory Meeting (8-10 October 2018)

October 2018

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/en/research/publications

Contact: library@ecmwf.int

European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, Berkshire RG2 9AX, England

Summary

This paper discusses the development and evaluation process followed at ECMWF to upgrade the Integrated Forecasting System (IFS), and illustrates how potential changes, developed and tested by individual scientists, are gradually merged and evaluated prior to their acceptance into the next version of the ECMWF IFS.

It discusses why, and how we are using a hierarchical testing strategy, whereby we test and merge changes stepwise and gradually. Individual changes are tested first in reduced configurations, then if accepted they are merged, and tested in more complete configurations. Finally, all the proposed changes that are deemed acceptable are merged and tested in all the IFS components at operational resolution. Evaluation relies on a use a range of metrics, selected to provide us with meaningful, statistical significant information on the impact of the cycle upgrade.

At the end of the process, a new cycle is judged to be ready for implementation if (a) it brings improvements to the ECMWF forecasts, and/or (b) it improves the realism of the simulated Earth-system, and/or (c) if it includes new forecast products that could improve our service to our users. All three aspects are considered when a final decision is taken. Thus, some model cycles could be implemented even if the impact on forecast quality is small or neutral, provided that new features that could lead to future improvements are implemented.

A focus of the current paper - indeed the primary driver for it - is to review whether the details of the hierarchical testing approach are optimal. Some changes are proposed in response to evolving strategic drivers and the increasing complexity of the full system.

1 Introduction

Final acceptance of a cycle for operational production follows the successful completion of an experimental suite (e-suite, designed to test all changes at operational resolution). In practice, acceptance (or otherwise) relies heavily on performance relative to the operational suite (o-suite), where performance is measured applying a range of metrics (headline scores, and a much wider set of scores and metrics covering many variables in a range of geographies and at different time ranges). Other considerations include fit to the ECMWF strategy and the potential of better science, even if not immediately beneficial in terms of scores, to be an enabler of future improvements.

E-suites are extremely expensive and can realistically only be completed once or twice a year. For the main cycles, they represent the culmination of a development and testing process that brings together many separate ideas and developments. It is not affordable to subject all these ideas to the full level of testing of an e-suite, so we have to rely on a hierarchy of different tests, appropriate to testing ideas at different levels of maturity. These span from the very first tests used by an individual scientist to see whether tentative ideas are worthy of further investigation, to tests used when bringing together various packages as we approach an e-suite.

A key consideration is the degree to which the cheaper tests in the hierarchy can be relied on to predict the results of the fuller more expensive tests that in an ideal world we would run; particularly where these tests are being used to make decisions (e.g. accepting that a change can pass to the next stage of testing, particularly if in that next stage it will be combined with other changes). The current approach is at least broadly successful, in as much as ECMWF's forecasts have continued to improve, and in so doing have maintained their world-leading position. However, for several reasons (detailed in Sections 1.1, 1.2 and 1.3) it is timely to review the model cycle evaluation process. This review follows the

appraisal of the 'Research-to-Operation' process, in 2017 (Buizza et al, 2017). These - and some of the specific questions they raise - are detailed below.

The remainder of this paper considers these issues, describes the process, proposes changes to our approach already thought to be required based on current evidence, and identifies areas where further work is required to determine the appropriate response.

More specifically, after this introduction, Section 2 outlines our testing strategy, which is based on a hierarchical approach that introduces in a controlled, step-wise manner all software changes, and illustrates the complexity of the operational suite. Then, in Section 3 we describe the evaluation process used to assess whether the new changes introduced in a new model cycle should be accepted. Section 4 is devoted to a discussion of ongoing work to improve further the model cycle testing and evaluation process, and how it could look in the future. Conclusions are drawn in Section 5.

1.1. Evolving strategic drivers

The ECMWF strategy 2016-2025 emphasizes the importance of ensemble predictions of high impact weather up to two weeks ahead, a seamless approach aiming at predictions of large-scale patterns and regime transition up to four weeks ahead, and global-scale anomalies up to a year ahead. New headline scores have been added (e.g. the frequency of poor near-surface temperatures in the ensemble) to the ones adopted a few years ago, to monitor progress of additional ECMWF products. The focus is no longer confined to the day-3-to-day-10 range, but goes from day 1 to year 1, with special attention devoted to the day-3-to-day-30 range.

These developments, and the increased focus on ensembles, shift the balance of considerations in the final decision as to whether a proposed new cycle is acceptable for operational adoption. In turn, this raises questions about whether the balance of considerations in earlier research testing carried out as part of the development of a new cycle remains appropriate. Much of the early testing has historically focused on the single, deterministic, high-resolution analysis and forecast (HRES), with testing of the ensemble of data assimilations (EDA) and the medium-range/monthly ensemble (ENS) performance often coming only late in the process.

In as much as improvements to the HRES typically carry through to improved ensemble performance [from the analysis to the seasonal ensembles], this might be appropriate. However, given the importance of the ensembles and that there will not always be a one-to-one correlation between HRES and ENS performance changes, we need to consider how to get earlier sight of the performance of the full ensemble system in a cost-effective manner. Similarly, with the desire to maintain a seamless system through the extended and seasonal time ranges (including the seasonal ensemble, SEAS), it is worth re-visiting our testing strategy to ensure an assessment of the performance at these longer lead times at the appropriate stage.

1.2. Issues arising from increasing complexity of the full system

The full system has grown in complexity, with, for example, all forecasts now featuring a coupled ocean (from 45r1, implemented on 5 June 2018) and the EDA being used to provide information to the 4D-Var analyses (Figure 1). While the introduction of such additional complexity has clearly led to

improved performance of the full system, it does bring into question whether the existing practices for cheaper research testing remain appropriate or whether they need to evolve.

For example, the standard 'workhorse' for early research testing has been the uncoupled TCo399L137 (spectral triangular truncation with 399 total wave numbers, with cubic-octahedral grid in physical space, and 137 vertical levels). Is there still a role for uncoupled testing now that all the forecast systems are coupled (and if so, in what circumstances)? Similarly, given the increasing importance of the EDA, what is realistic in terms of costs to get earlier sight of likely performance problems, whether in the EDA itself or going further and looking at the feedback between the EDA and 4D-Var (either through changing background error co-variances or errors of the day)?

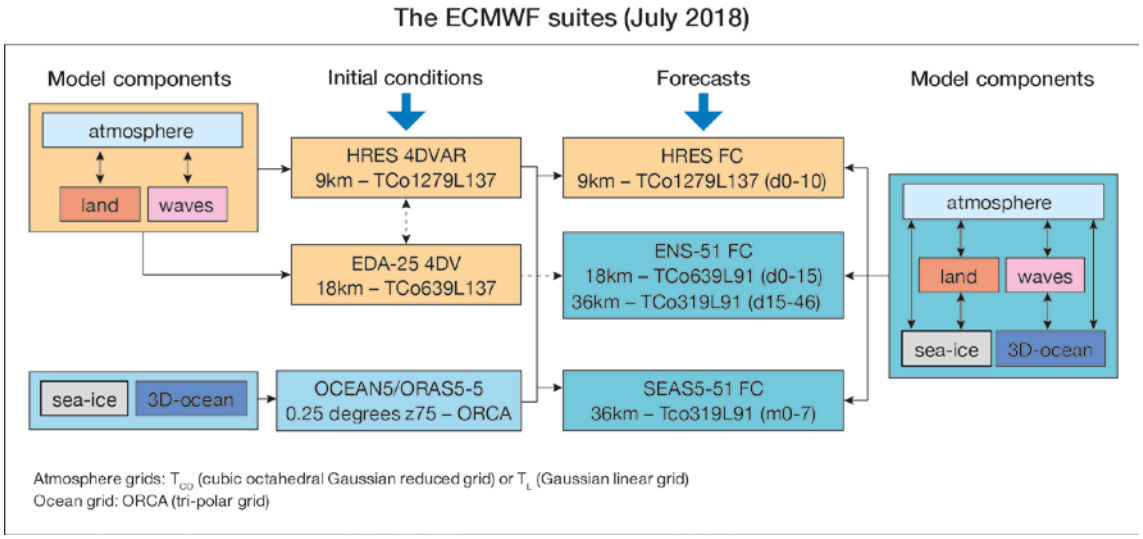

Figure 1: Schematic of the components of the ECMWF operational suite (see text). The blocks in the second column illustrate the fact that the initial conditions are generated using the high-resolution 4-dimensional variational assimilation (HRES 4DVAR) and the Ensemble of Data Assimilations (EDA) for the atmosphere/land, and the NEMOVAR analysis (OCEAN5/ORAS5) for the ocean and the sea-ice. The blocks in the first column highlight the model components used in the analysis: note that the 4DVAR/EDA and the NEMOVAR analyses run independently. The blocks in the third column list the main characteristics of ECMWF's three operational forecasts, the single high-resolution (HRES), the medium-range/monthly ensemble (ENS) and the seasonal ensemble (SEAS5). The blocks in the fourth column highlight the fact that all forecasts are run with a coupled atmosphere, land, waves and ocean (sea-ice and 3D ocean) model. The dashed-arrow linking the EDA and the 4DVAR blocks indicate that the EDA is used in the 4DVAR analysis to compute flow-dependent background error statistics. The dashed-arrow linking the EDA and the ENS blocks indicate the fact that the EDA is used to define the ENS initial perturbations. Note that, for simplicity, the background forecast used in the HRES 4DVAR assimilation is not shown in this figure: it is worth to point out that this (coupled) background forecast is initialised by the OCEAN5 sea-ice, so the assimilation is based on a weakly-coupled sea-ice/atmosphere assimilation approach.

1.3. Responding to lessons learned

Additionally, as part of a process of continuous improvement, it is appropriate to consider any issues encountered in recent cycle implementations and assess whether any changes in approach are required. For example, some evidence from 45r1 suggests that we might not have a consistent view of how to interpret verification scores against analyses; particularly for developments that change the characteristics of the analyses. What are our metrics to decide whether these are positive? Is the

infrastructure in place to enable us to look confidently and easily at the scores and diagnostics that we should be considering - or is availability of the tools itself a limitation?

2. The ECMWF testing hierarchy, cost considerations and technical set-up

This section describes the Integrated Forecast System (IFS), its main components and how they interact. It documents the cost of the various components of the operational suite (o-suite), describes the Research to Operations (R2O; Buizza et al, 2017) process and the testing hierarchy, and provides details and costs of the simplified configurations used in research experiments. Finally, it discusses the efficiency of the testing and R2O processes.

Forecast component	Description	#	Atmos. horiz. and vert. res.	Forecast length	3D ocean and sea-ice
4DVAR	Atm/land/waves High-resolution analysis	1	9 km 137 levels	-	-
EDA[25]	Atm/land/waves Ensemble of data assimilations	25	18 km 137 levels	-	-
OCEAN5[5] **/ ORAS5**[5]	3D ocean and sea-ice Ensemble of analyses	5	-	-	25 km 75 levels
HRES	Atm/land/waves/3D ocean High-resolution	1	9 km 137 levels	0-10 d	25 km 75 levels
ENS[51]	Atm/land/wave/3D ocean Medium-range/monthly ensemble	51	18 km 91 levels	0-15 d	25 km 75 levels
			36 km 91 levels	15-46 d	
SEAS5[51]	Atm/land/waves/3D ocean Seasonal ensemble	51	36 km 91 levels	0-7 m (0-13 m quarter)	25 km 75 levels

Table 1: The 6 components of the ECMWF Integrated Forecasting System (IFS) at the time of writing (July 2018): description (column 2), number of ensemble members (column 3), resolution of the atmosphere/land components (column 4), forecast length (column 5) and coupled ocean/sea-ice model and resolution of the ocean model (column 6). Number subscripts indicate that the component is run in ensemble mode, and indicates the number of ensemble members. For the ocean, OCEAN5 denotes the daily analysis cycle suite and ORAS5 the delayed cycle (see Mogensen et al 2012a,b for details).

2.1. The IFS operational suite

The IFS consists of the following primary components (see Table 1):

- Ensemble of data assimilations;
- High-resolution single analysis and forecast;
- Medium-range/monthly ensemble forecast;

- Seasonal ensemble forecast;
- Ocean ensemble of analyses.

Figure 2 shows the utilisation of the operational HPC (node count vs. time of day): note that every day during the production hours, the operational suite (dark green) uses up to about 85% of the usable nodes of one cluster. Figure 3 shows a Gantt chart of the daily schedule showing when components run, and how they link together and depend on each other: its complexity is evident.

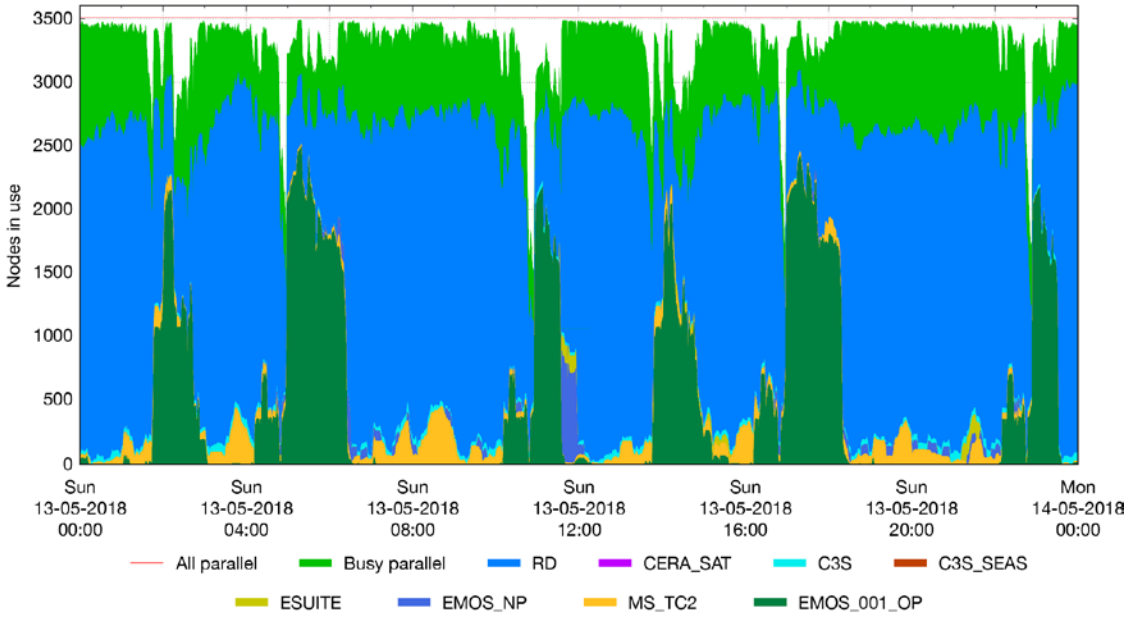

Figure 2: Utilization of the operational HPC cluster (node count vs. time of day) for one calendar day. Among them, the dark green areas, which highlight the ECMWF time-critical production runs, the light blue areas which identifies the research experimentation, and the yellow areas the Member States' time critical applications.

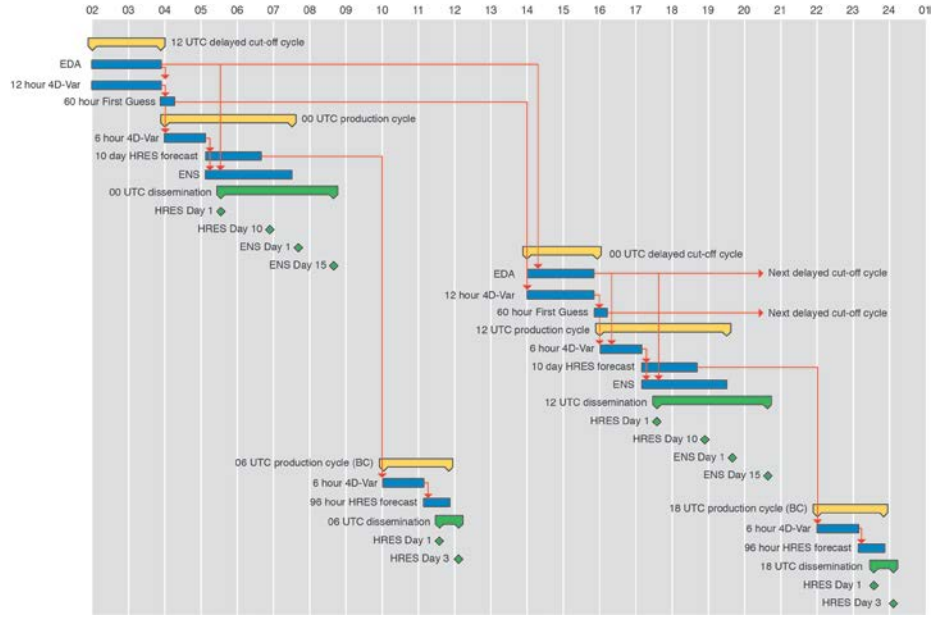

Figure 3: Gantt chart of the daily schedule showing when components run, and how they link together and depend on each other.

2.1.1. EDA

The EDA (Buizza et al, 2008; Isaksen et al, 2011; Bonavita et al, 2017) comprises 25 members, each running a TCo639L137 resolution (TCo stands for spectral triangular truncation with cubic-octahedral grid). Each analysis is performed with a 4D-Var system, with a 12-hour time window, and two minimizations with resolution TL191/TL191 (TL stands for spectral triangular truncation with linear grid in physical space).

2.1.2. HRES

The single high-resolution analysis (Rabier et al, 2000; Bonavita et al, 2017) and forecast (Wedi et al, 2015) includes a one-member analysis and a one-member, 10-day forecast, both with a TCo1279L137 resolution, which is equivalent to a grid resolution of about 9 km. At 00 and 12 UTC, ECMWF produces an early-delivery, 6-hour window analysis, and a 12-hour, long-window analysis (LWDA). These analyses are generated with the same 4D-Var system as the EDA, but with three minimizations (i.e. one more than in each EDA member) and higher resolutions (TL255/TL255/TL399). The early delivery high-resolution analysis is also generated at 06 and 18 UTC, as part of the Boundary Condition project. Since cycle 45r1 the HRES forecast has been coupled to the NEMO ocean model. Please note that, hereafter, we use the term 'HRES' to mean the single high-resolution analysis and forecast configuration, as opposed to the ensemble ENS.

2.1.3. ENS

The medium-range/monthly ensemble (Buizza and Palmer, 1995; Molteni et al, 1996; Vitart et al, 2014; Leutbecher, 2018) has 51 members, run with a TCo639L91 resolution, which is equivalent to about 18 km in physical space, up to forecast-day 15. ENS is extended twice weekly (at 00UTC each Monday and Thursday) to 46 days, with a TCo319L91 resolution (about 36 km).

2.1.4. OCEAN5/ORAS5

OCEAN5 is the fifth generation of ocean analysis system at ECMWF (Zuo et al, 2018). It comprises a behind-real-time (BRT) component, that was used for production of ORAS5 (Ocean Re-analysis 5); and a real-time (RT) component, that is used for generating daily ocean analyses for NWP applications.

OCEAN5 uses the NEMO (Nucleus for European Modelling of the Ocean; Madec, 2008) ocean model in a global configuration ORCA025.L75 (Bernard et al, 2006), which is coupled to the LIM2 (Louvain-la-Neuve sea-ice model version 2, see Fichefet and Maqueda, 1997) sea-ice model. The data assimilation system used in OCEAN5 is NEMOVAR (Mogensen et al, 2012a, 2012b) in its 3D-Var configuration using the first-guess at appropriate time (FGAT) approach. The OCEAN5 BRT and RT components are linked to each other. The BRT uses 5-day assimilation window and is updated every 5 days, with a delay of 7-11 days to allow ingesting more observations. The RT analysis updates daily using a variable assimilation window of 8-12 days, and is always initialized from the last day of the BRT analysis.

2.1.5. Operational computing cost of the IFS components

Table 2 gives a summary of the costs of the various components, in terms of real-time duration of production, the number of computer nodes, and the frequency they are run every day.

Component	Duration	Node count	Run frequency	Total node-hours per day
LWDA HRES Analysis (the 12-hour, long-window analysis; run at 00 and 12UTC)	1:13	704	2/day	307
LWDA HRES Forecast (run only up to 60h from the LWDA analysis)	0:24	352	2/day	282
EDA Analyses	1:02	1820	2/day	3761
EDA Short Forecasts	0:13	1092	2/day	473
EDA Jb Calculation	0:35	336	2/day	392
HRES Analysis (the 6-hour window early-delivery analysis, run at 00 and 12)	0:46	704	2/day	540
HRES Forecast (the 240h forecast)	1:05	352	2/day	763
Product Generation	1:21	300	2/day	810
ENS 'Chunk 1' Forecast (fc day 0 to 15)	0:55	1632	2/day	2992
ENS 'Chunk 1' Product Generation	0:49	168	2/day	274
ENS 'Chunk 2' Forecast	0:42	1632	2/day	2285
ENS 'Chunk 2' Product Generation	0:21	168	2/day	118
ENS Monthly Extension	0:55	765	2/week	200

Table 2: Summary of the costs of various components of the operational suite, in terms of real-time duration of production, number of computer nodes, frequency they are run, and average total node-hours per day. Apart for the usual acronyms already explained in the text, in column 1 LWDA indicates the long-window (12-hour) analysis run at 00 and 12 UTC, and 'HRES analysis' indicates the early-delivery, 6-hour window analysis run at 00 and 12 UTC. Please also note that the 51-member ENS is run in 4 chunks, two to generate the forecasts and two to generate the products. Please note that the 06 and 18 UTC Boundary Condition analyses and forecasts are not listed in this table.

Note that, in addition to the components listed, the 06UTC and 18UTC boundary-condition runs, which are not shown, represent a substantial use of compute resources. Furthermore, there are very many smaller tasks in the production suites that are individually too minor to list explicitly, such as observations processing in the analysis-type suites, EDA error calculation, standalone wave model runs, singular vector calculations in the ENS suite, and other general post-processing tasks. These also represent a substantial additional use of computer resources.

Even with these omissions, the production runs are very costly, and represent nearly 20% of the total node-hours available in a single cluster. Moreover, because of the highly peaked nature of the production runs, which is clear from Figure 2, the instantaneous usage is often much higher than the average of

20%; during the so-called 'model hour', when both the ENS forecasts and the HRES forecasts are running, along with their product generation suites, about 85% of the usable nodes of one cluster are in use at the peak (since about 15 years ago, the ECMWF High Performing Computer includes two identical clusters). This is the reason why the current resolution is the highest possible today, given that we want to be able to complete the operational production in the same amount of time even in case the operational computer cluster has problems, and we need to use the back-up one. This also means that testing IFS cycle upgrades at the full operational resolution must be planned carefully, and can only be done with very few configurations (see discussion in the following sections).

2.2. The medium- /extended-range testing hierarchy

New versions of the IFS (cycles) are released once or twice per year; the operational implementation of each cycle is the culmination of a 9-12-month R2O process. The exact timing of this process depends mainly on the complexity of the new cycle's upgrades, on how smooth the merging and testing proceed, on the availability of computing power to complete all planned testing, and on the length of period during which the o- and the e-suites are run in parallel (this ranges to a minimum of 4 weeks for upgrades that do not involve changes in resolution, and at least about 3 months for changes that include them).

During the R2O process, we progress through a series of phases and stages of increasing integration, and a hierarchy of tests of increasing complexity:

- The Alpha-phase testing, which includes five stages:
  - 0 - Individual ad-hoc testing
  - 1a - Individual testing against controls
  - 1b – Thematic pre-merging and addressing interactions
  - 2 - Incremental build and testing
  - 3 - RD e-suite: higher resolution HRES/ENS and EDA testing
- The Beta-phase testing;
- The Release-Candidate-phase testing.

2.2.1. Alpha-phase, Stage 0 - Individual ad hoc testing

Before the beginning of the process, researchers are working on a portfolio of new IFS developments, individually or in small teams. In general, these contributions are independent of each other, and segregated in different branches of the version-control repository, so are developed in isolation.

At this point testing is ad-hoc, according to the requirements of the contribution, and not centrally managed. Researchers run experiments with their branches on the ECMWF High Performance Computing (HPC) facilities using the PrepIFS submission system.

As part of this testing, individual developers assess the impact on the computing cost of the proposed changes, and document them together with the expected impact on performance. If the proposed changes have a negative impact on computing cost that was not expected, the developers are asked either to find

ways to make them cost neutral, or to make the case to the evaluation board that the changes must be included, even at the unplanned, extra cost.

A recent introduction, in Autumn 2017, is a system that allows developers to build the full IFS locally, along with all its dependencies, and run a small suite of model tests, to ensure a basic level of technical correctness much more quickly than was previously possible. The model tests are performed at very low resolution, and run only a few time-steps, to ensure they complete in a few seconds each. The virtue of this is that potentially very many tests can be run, exercising a wide variety of configurations and therefore covering a large proportion of the source code. Also, it is potentially easy to run tests across a range of different platforms, compilers and compiler options. Initially, a dozen tests were included, which cover basic forecast, tangent linear and adjoint configurations, parallel test, uncertainty tests, climate tests, and atmospheric composition configurations. Tests can be run using the GNU compiler on individual Linux workstations. Developers have been encouraged to provide their own tests alongside new IFS features or bug fixes; only the first five of the 12 tests were provided by staff developing the test system, the remainder being introduced by researchers wanting to use the system. Cy46r1 saw the introduction of a further six tests, taking the total to 18.

Key improvements in this stage of testing will be to extend the number and diversity of tests, and the range of compilers, compiler options and platforms that the system can be run on, as well as increasing the uptake of the system amongst researchers. Further improvements could include adding tests to the suite that represent Météo-France or ALADIN/HIRLAM configurations.

2.2.2. Alpha-phase, Stage 1a - Individual testing against controls

Stage 1 of the process formally begins when the previous cycle is declared, which is to say that its content is frozen. (Note this is still several months before that cycle goes into production.) At this point, ‘official’ control experiments for the next cycle are created centrally, which other researchers can use for the purposes of comparison. Four control experiments are created: both HRES and ENS controls, for both the most recent summer and the most recent winter seasons. (Note that the creation of ENS controls is a relatively recent addition, for Cy46r1 cycle development; historically only HRES controls were provided.)

Once these controls are available, researchers can test their model changes, and compare them with the control experiments. The controls are, as standard, at reduced resolution: TCo399L137 for HRES and TCo399L91 for ENS. Harmonising research testing at TCo399 horizontal resolution was agreed prior to the development of IFS cycle 43r3, after a thorough study that balanced representativeness against computational cost. As part of this testing, the cost of the planned changes is re-evaluated against the controls.

To reduce the cost of ENS tests, only 21 members are run, and forecasts are initialised from operational analyses every eight days (rather than twice per day) over the course of one year; winter/summer experiments coupled to own-analyses should be a future addition (see Section 4).

The HRES controls provided for Cy46r1 development were for the first time coupled to the ocean, to reflect the fact that since Cy45r1 all ECMWF forecasts are generated with a coupled ocean, land, sea-ice and atmospheric model. Note that the ocean resolution in the TCo399 tests has not been scaled down from the operational value of 0.25°; therefore, at this resolution the NEMO fraction of the forecast

compute time is about 50% of the total; a much higher proportion than in production runs. Accordingly, work is ongoing to evaluate the suitability of a 1° ocean for testing (see Section 4).

2.2.3. Alpha-phase, Stage 1b – Thematic pre-merge testing

After stage 1a an initial evaluation of all changes is made, and a selection is agreed for merged thematic tests. This reduces perhaps 100 individual tests to 10 or less to be provided for Stage 2. The goal is to identify, at an early stage, contributions that interact with each other meteorologically, and to assess the impact of the thematic-merged changes on computing cost. Typically testing continues at the TCo399 research resolution, and comparisons with control experiments are made. In case a change is expected to have a specific effect at the operational resolution, and/or a substantial impact on computing cost (say leading to a cost increase of more than ~ 2%) the testing is done also at full resolution. In case computing cost at full resolution increases in an unplanned, substantial way, the contributions leading to extra costs are re-assessed, and are accepted only after further cost-benefit evaluations.

Stages 1a and 1b usually last about two months.

2.2.4. Alpha-phase, Stage 2 - Incremental build and testing

Stage 1 ends and stage 2 begins with the cycle handover deadline, from which point the new cycle is built incrementally in a series of versions of increasing completeness. Each version is tested against both the summer and winter HRES and ENS control experiments. The computing cost of the new cycle, compared to operations, is also confirmed.

During this phase, one or two scientific co-ordinators are appointed to lead the analysis and verification of the new cycle. The Implementation Working Group (IWG) responsible for the implementation is convened and meets regularly to monitor results and assess progress.

From Cy45r1 onwards, scientific ENS testing has been conducted for the incremental builds, consisting of TCo399/21-member runs initialised from an operational analysis, every 8 days for 1 year. Starting with Cy46r1, a second test stream was added, in which each experiment is initialised from the corresponding analysis experiment, i.e. using the same incremental merge version. This enables us to more easily compare the long forecast and the ensemble directly, since they will start from the same initial analysis.

Stage 2 usually takes about 2 months.

2.2.5. Alpha-phase, Stage 3 - RD e-suite

After the incremental merge of the new cycle has reached a final candidate version, which is considered acceptable by the IWG based on the low-resolution testing in stage 2, testing at high resolution begins. This means the TCo1279L137 HRES and TCo639L91 (51-member) ENS operational configurations. The EDA is also tested at this stage - this means that the EDA background error statistics from the new cycle can be used in the high-resolution testing.

Historically, the final stage involved only the EDA/HRES. The high-resolution ENS testing was added to the formal process with Cy45r1. These runs are initialized from the HRES analysis experiment and EDA experiment (discussed below), with one forecast per day over each of the summer and winter

periods. The TCo639 ENS experiment runs faster than the HRES analysis, allowing one ENS forecast per day, following each of the high-resolution analyses.

Full-resolution testing is conducted only at this relatively late stage due to the high cost of these configurations, although this creates a vulnerability to problems that manifest only at high resolution. Naturally, problems found in the later stages of merging can be more difficult to identify and isolate than if they were found earlier. By the time high-resolution testing begins, of the order of 100 independent contributions of varying complexity have been merged together.

It should be noted that some ENS products (e.g. the Extreme Forecast Index), need reforecasts (i.e. forecasts with past initial conditions but re-run with the new candidate operational configuration) to evaluate the model climate. Thus, an integral part of the testing strategy is to assess the impact of IFS changes on the reforecasts (see Appendix A). For the ENS monthly extension, the reforecast record covers the last 20 years, while for seasonal forecasting the reforecasting period needs to be as long as possible, and it usually starts from 1981 (1981-2016 for SEAS5; Stockdale et al, 2018). Reforecasts are initialized from reanalyses (atmosphere, ocean, land and sea-ice).

The skill of the monthly forecast range must be evaluated for each new operational implementation. The evaluation is typically done with a reduced set of re-forecast and ensemble members, considered sufficient to assess the skill in predicting the MJO and its mid-latitude teleconnections, evaluate the model biases, and obtain statistically-significant values on a set of scores representing the skill at weeks 1-4. The research conducted under the WWRP/WCRP Subseasonal-to-seasonal (S2S) programme is yielding new insights into predictability at the extended range, and should in the future be included in the evaluation of model cycles.

Stage 3 takes about 2 months.

2.2.6. Beta-phase testing

The testing described thus far is conducted in ‘research mode’: that is, using standalone experiments run by RD scientists using the PrepIFS environment. Once the full-resolution testing previously described has been evaluated, the new IFS Cycle is declared, at which point the transition from research into operational mode begins. This closes the ‘Alpha phase’, which is followed by two further phases.

Once a cycle has passed the initial Alpha-testing phase, it is subject to further tests to evaluate the potential impact on Production. In this phase, an “e-suite” (experimental suite) using the new cycle runs in parallel with the current operational “o-suite”. A technical test data set is made available to users to test the technical impact on users’ applications and tools. This testing phase is usually prone to minor updates and restarts and users are asked not to rely on the availability of these test data sets for their own tests before the next stage has been reached. During this phase, there might be a few changes that alter the model results, but these are fully documented. Users have access to only part of the data generated.

As a prerequisite to successful Beta-phase testing, a significant amount of effort goes into first ensuring the e-suite is bit-identical to the high-resolution research experiment. The research experiment also serves as a warm-up phase for the e-suite. This is a very important feature in terms of model error and variational bias cycling.

The Beta-phase testing lasts about 2 months.

2.2.7. Release-candidate-phase testing

For the final testing phase, the suite is frozen and runs at or close to real-time. Changes that alter the model results are not permitted. A full set of product services (e.g. dissemination of test data, ecCharts, etc.) are offered during the whole period. Users have access to all data generated during this phase.

The release-candidate-phase lasts one month for normal changes and three months before the planned implementation of a high-impact change (such as resolution increases for which users are given access for a longer time).

2.3. Computational cost of Research Department (PrepIFS-based) experiments

The HPC resources used for a given experiment type and a given resolution are pre-defined and carefully tuned with each new HPC platform, and potentially with new cycles, to balance throughput and time-to-solution.

Table 3 gives the costs of the different types of experiment, in comparison with their operational equivalent. Even where the resolution of the research experiment is the same as that of the operational component, the research version is typically cheaper, since the time-to-solution requirement is greatly relaxed.

Alpha Stage	Research experiment type	Node-hours per day (A)	Node-hours per day of equiv. oper. component (B) (from Table 2)	Percentage of equivalent operational component (A/B)
0-2	TCo399 HRES 2×(4D-Var + 10d fc) + SEKF	88	540 + 763	7%
0-2	TCo399 21-member ENS 2×(15-day forecast)	485	2992 + 2285	9%
3	TCo1279 HRES	759	540 + 763	58%
3	TCo639 51-member ENS	4838	2992 + 2285	92%

Table 3: Cost (expressed in terms of node-hours per day of equivalent operational component) of the experiments run in RD during the alpha testing of the single, high-resolution configuration (HRES) and the medium-range ensemble (ENS). [The TCo639 51-member ENS cost was estimated as (51/21)×(639/399)3 of the cost of the TCo399 21-member experiment.] Note that the table does not include the cost of testing on the seasonal time scale: see Appendix A for a discussion of these costs.

2.4. The Copernicus Atmospheric Monitoring Service (CAMS) suite

The atmospheric composition forecasts run as a separate production suite in parallel with the medium-range forecast suite (using a lower spatial resolution to compensate for the additional cost of the atmospheric chemistry and composition calculations). Although IFS cycle upgrades for CAMS do not need to occur at the same time as for the medium-range, an effort has recently been made to ensure the upgrades are coordinated, so that the two suites are always using the same cycle. This was first done with cycle 45r1. As a result, the development and testing phases for composition and NWP are synchronised, leading to large reduction in effort.

CAMS researchers are involved throughout the IFS cycle upgrade process, contributing improvements to each cycle, and testing the incremental merges (Alpha testing, stage 2) in a CAMS test configuration.

Certain CAMS forecast configurations are already included in the low-resolution, quality assurance test suite described earlier (Alpha testing, stages 0-1). The process of testing the incremental builds in a CAMS configuration against relevant controls is not centrally managed in the same way as for the HRES and ENS configurations, but mirrors the process and is run in parallel by CAMS researchers. Usually, certain incremental build stages have no impact in the CAMS context (particularly those relating to spinning-up a new EDA, since CAMS uses a static covariance matrix). Hence not every stage is tested as rigorously as for HRES and ENS. Typically CAMS technical tests will begin with the build that includes the CAMS contributions themselves, and full scientific testing is done once any dynamics, physics or assimilation changes that are likely to impact CAMS have been merged.

The primary scientific evaluation of CAMS is done externally under contract, both for each CAMS e-suite prior to implementation, and of the o-suite on a quarterly basis. This evaluation is outside the scope of this paper.

2.5. Seasonal forecast system

The seasonal configuration is run separately from ENS, as another separate suite. Upgrade of the seasonal system occurs only about every five years, during which time several IFS Cycles are released and go into production for the medium-range/monthly time scale and CAMS. For the cycles that are 'skipped', a full evaluation of seasonal forecasts is not performed, due to the high computational cost. As for the ENS, SEAS5 needs a re-forecast data set. The re-forecasts are used for product calibration and for skill assessment.

Although the atmospheric resolution for seasonal ranges is lower than for the medium-/extended-range, the need for re-forecasts implies that a full evaluation of these systems is expensive. For the evaluation of developments and cycles prior to implementation, a reduced re-forecast set is chosen. Because of the expense, only those developments thought to affect the seasonal forecast skill are tested. Thus, developments in the coupled model and the ensemble generation need to be evaluated, but changes in the atmospheric assimilation system are usually not, since as the lead time increases, the influence of the atmospheric initial conditions in the forecast decreases. However, developments in the analysis of the 'slow' components (ocean, land and sea-ice) do need to be evaluated, especially those affecting the consistency between the re-analyses and the real-time initial conditions. The impact of new re-analysis products also needs evaluation. Appendix A has a more detailed discussion of the testing strategy for the re-forecast suite.

2.6. Testing and evaluation of ocean and sea-ice component

The ocean component of the IFS is updated only every few years. Changes in the ocean model (NEMO) and its data assimilation component (NEMOVAR), and in the observational data sets, are linked to the production of new ocean reanalyses/analyses, which provide ocean initial conditions. To the time needed for implementing and testing ocean updates we must add the ocean re-analysis production time, which is carried out sequentially. It took approximately 10 months to produce the current ocean reanalysis (ORAS5) for the period 1975 to 2015.

We take as an example the most recent update to illustrate the process of upgrading the operational ocean component. This included the upgrade of the ocean model version from v3.0 to v3.4, increases in the horizontal and vertical resolution, inclusion of a dynamical sea-ice model, upgrades in the

NEMOVAR ocean data assimilation, and revised coupling with the wave model. From June 2018, OCEAN5 also initializes the ocean component of the coupled HRES.

The evaluation of ocean model and ocean re-analysis performance involves the production of multi-year integrations. Fit to observations, error growth, mean state, and variability are evaluated using observational data sets, to form a standard set of objective diagnostics. There is also subjective expert evaluation. For example, the evaluation of assimilation of sea-level from altimeter data accounts for plausible values of trends in global steric component.

Data assimilation is also tested in short integrations, looking at fit to observations and error growth. When changes involve the modification of the background error covariance formulation, a set of single observation experiments is also conducted.

The low-resolution configuration with a 1° horizontal resolution and 42 vertical levels (ORCA1_Z42 grid) has been extensively used for testing ocean model and data assimilation developments. This configuration was used in OCEAN4; it is much cheaper than the 0.25° horizontal resolution and 75 vertical levels (ORCA025_Z75 grid), and produces multi-year time series at a much faster rate. In the future, we plan to use a new configuration with a 1° horizontal resolution and 75 vertical levels (ORCA1_Z75 grid) configuration as a cheap surrogate of ORCA025_Z75, since both share the same vertical resolution, an important aspect for testing vertical physics.

3. Evaluating ECMWF analyses and forecasts

Assessing the quality of forecasts for both the current operational system and proposed changes through the IFS Cycle development process is a constantly evolving activity. There are increasing numbers and types of output products, changing user requirements, tighter constraints from additional observations, and increasing emphasis on wider aspects of the Earth system and forecast time ranges. The direction of development of the IFS evaluation process is closely linked with the ECMWF Strategy and includes:

- Improving evaluation methods and tools with a priority on efficient use of resources;
- Using a wider range of observations from different sources for a more holistic evaluation;
- Improving the evaluation of near-surface and high-impact weather;
- Increasing emphasis on the evaluation of ensembles of analyses and forecasts through to extended-range timescales;
- Extending evaluation methods and tools for the atmosphere to other parts of the Earth system (ocean, land, sea-ice, atmospheric composition).

This section describes relevant aspects of the evaluation process. First, we describe some general principles for measuring IFS performance; what aspects are most important and how do we measure “improvement” (section 3.1). Second, we describe the current process, software tools and metrics to evaluate the changes for each IFS cycle (section 3.2). Thirdly, as part of a process of continuous improvement, it is important that the evaluation process not only assesses the impact of changes for an IFS Cycle, but also provides information that can feed through to future developments (section 3.3). The reader is referred to Section 4.3 for a discussion about areas of future developments of the evaluation process.

3.1. Some principles for measuring improvements to the IFS

A key activity at ECMWF is to improve the forecasts for users. In assessing the impact of changes for a new Cycle, it is therefore necessary to determine how to measure improvement. There are different aspects to consider and the decision on whether to accept changes will always be a judgement call balancing many different considerations. However, there are several principles that guide the decision process, as illustrated hereafter.

3.1.1. Holistic evaluation

It is important to have as holistic a view as possible to evaluate the impacts of changes to the system. This requires a comprehensive assessment covering many different aspects:

- Headline scores for the operational forecasts;
- Quantities representative of the evolution of the large-scale flow (e.g. geopotential height, upper air temperature and winds) and near-surface weather (e.g. 2m temperature, 10m winds, precipitation, cloud cover, surface radiation, and user-driven products such as pseudo-satellite images).
- Quantities representative of high-impact weather (e.g. heavy precipitation, precipitation type, lightning, …) and metrics for tropical cyclone evaluation;
- Metrics for large-scale meteorological phenomena (e.g. blocking frequency, NAO-index, ENSO, MJO) and teleconnections;
- An assessment of forecasts against both analyses and observations for a range of forecast lead times: from the analysis and 12-hour first guess used in the assimilation, through the medium-range to the extended and seasonal timescales more representative of the model climate;
- Various resolutions and time-steps used in different configurations (HRES, ENS and SEAS);
- Different geographical regions and seasons (a minimum of one winter and one summer, but spring/autumn where required, e.g. changes to the spring snow melt);
- A variety of metrics that measure the amplitude of the error (e.g. bias, standard deviation, root mean square error), the pattern of the error (e.g. anomaly correlation), the activity in analysis and forecast, categorical scores (e.g. SEEPS for precipitation) and metrics for the probabilistic skill of ensemble forecasts (e.g. CRPS, ETS, EFI skill score);
- Technical evaluation (computational cost, memory usage, code refactoring).

Clearly, the more comprehensive the assessment, the higher the use of resources and larger the volume of information. So, a judgment is always required on the priorities and scope of the assessment depending on the specifics of the change and the expected impacts.

3.1.2. The importance of statistical significance

Every change to the IFS should be evaluated with consideration of statistical significance. Both the Earth system and the models that represent it are subject to rapid chaotic error growth, meaning that even the

smallest numerical perturbation to the forecasting system can lead to apparent changes in forecast skill and model climate that in fact are just a result of chaotic variability.

Recent work (e.g. Geer, 2015) has highlighted several aspects that need to be taken into consideration. Although sometimes large changes with systematic impacts can attain statistical significance rather quickly, often long time-periods are needed: as much as 6 months for typical small (e.g. 0.5%) changes to atmospheric medium-range scores and sometimes longer for regional evaluation. Secondly, in assessing statistical significance it is important to consider the time-correlations of forecast error that inflate the true uncertainty range, and to address this with increasingly sophisticated correlation models (e.g. AR1 and AR2). Finally, there is a need to correct for statistical multiplicity to inflate the error bars when comparing many different experiments, such as is typical in the R2O process.

There is a very careful balance to be struck between accepting "false positives" (e.g. false signals that a change has degraded or improved forecast scores) while still having some "power" in the statistics to resolve differences between versions. Because statistical significance is harder to attain further in to the forecast there sometimes has to be a judgement based on the first few days of the forecast and short-range metrics, such as the first-guess fit to observations, that attain statistical significance on much shorter timescales.

The above aspects of statistical significance are all included when assessing the impact of changes to the IFS for various quantities. Experiments during the initial testing stages may cover a month or two. Standard RD tests for thematic merges and later testing stages generally cover 3 months of summer and 3 months of winter which combined give a 6-month testing period.

3.1.3. Own-analysis and observation-based verification

To assess whether a forecast state is improved, we need to know the true state at the verification time. However, we can only ever have an approximate knowledge of the true state. As the forecast has improved over time, the magnitude of the forecast error has become closer to the magnitude of observation and analysis uncertainty, so evaluation needs to be interpreted with increasing care.

In experiments which make smaller changes to the analysis than the forecast, own analysis verification is a good metric because the analysis is the most accurate estimate of “truth” available, and is available everywhere. However, when there are changes to the IFS that affect the relative weights given to the observations and the background in the assimilation system, evaluation of the short-range forecast against own-analysis is no longer a reliable measure of forecast skill, leading to either apparent false positive or false negative impact on scores. Such changes include direct adjustments to specified observation errors, changes to the EDA or modifications to climatological background errors, for example associated with increased vertical or horizontal resolution.

If the analysis is drawn closer to the observations (further from the background), then the short-range forecast can be a poorer fit to the analysis than before, even though the analysis may be more realistic and improve the forecast at longer lead times. Hence, in these cases, what appears to be (possibly even a large) degradation (or improvement) in the early part of the forecast range (first couple of days in the extra-tropics but sometimes longer-range in the tropics) can be just an artefact of changes to the analysis. Therefore, in these cases, own analysis verification alone cannot lead to firm conclusions about the validity of a change and evaluation against independent analyses, or observation-based verification, are

essential. There is no definitive way to determine whether an analysis is better or not in these circumstances, as all analyses and observations have uncertainties, but evaluation against a wide ranging and consistent set of observations, evaluation against the current operational analysis and evaluation of the impact on the medium-range forecasts are all necessary to give confidence in the changes. Other centres often verify against ECMWF's analysis and there may be value in ECMWF verifying against an independent analysis (e.g. Met Office), but this is not currently done.

Evaluation against observations is also subject to difficulties. Whereas the analysis is global, observations are not distributed uniformly around the globe. Particularly, surface station reports (SYNOP) and radiosondes are clustered, with higher density in populated land regions and very little coverage over the oceans, high latitudes and large areas of the Tropics. An evaluation against observations can therefore give different results from an evaluation against analysis because they have weighted regions differently. This was seen in 45r1 where ocean wave validation against buoys and satellite altimeter data appeared to give contradictory results, but in fact this was down to the geographical distribution of the buoy data. In general validation against satellite data is under-utilised but, given the global coverage, offers many advantages over validation against surface in situ observations.

A second difficulty is uncertainty in the observations themselves. As was shown during the evaluation of the ENS 45r1 e-suite, considering observation uncertainty in the upper-air verification against observations (radiosonde data) can change the magnitude and even the sign of the evaluation differences (e-suite compared to o-suite). The CRPS is minimized for a reliable forecast only if observation uncertainty is considered. To do this more generally and more regularly, we would have to find appropriate observation uncertainty estimates for a wider range of quantities which are not routinely assimilated.

3.1.4. Communicating and decision-making: a focus on what's important

Any evaluation provides a large amount of information, and there will always be a need to summarize this in a form that can be readily and quickly understood. However, different target groups have different priorities. For internal IFS development and acceptance of changes, it is necessary to have a wide range of information available in different forms such as time-averages of scores with lead time, cross-sections, maps, and comprehensive scorecards, to condense the performance information on large-scale and near-surface weather parameters.

For forecast users, the focus is on the impacts that the forecaster is likely to notice from day to day, such as the changes to the characteristics of 2m temperature or precipitation, or the availability of new products such as lightning (introduced in cycle 45r1).

For other stakeholders, it is important to have a condensed summary for the main developments and representative measures of the forecast performance: for example, focused scorecards and time-series of headline scores. Summarizing the information from a complex system in a single number with weights for different components, such as for an "NWP index", is not considered at the current time. Although it can be useful in certain circumstances, it should be used with care as it can skew priorities and hide the complexities of the performance changes. To date, the synthesis provided by the ECMWF scorecards (see Section 3.2.2) is thought to be a good compromise.

3.1.5. A long-term perspective

It is recognized that not everything can be improved or be neutral in each cycle. There is usually a need to accept some degradation in certain aspects of the evaluation; development would stagnate if we had to achieve improvements across all performance measures. The forecast has many compensating errors, and some changes that are scientifically justified do not lead to immediate improvement in every aspect, but are a necessary part of the longer-term development process.

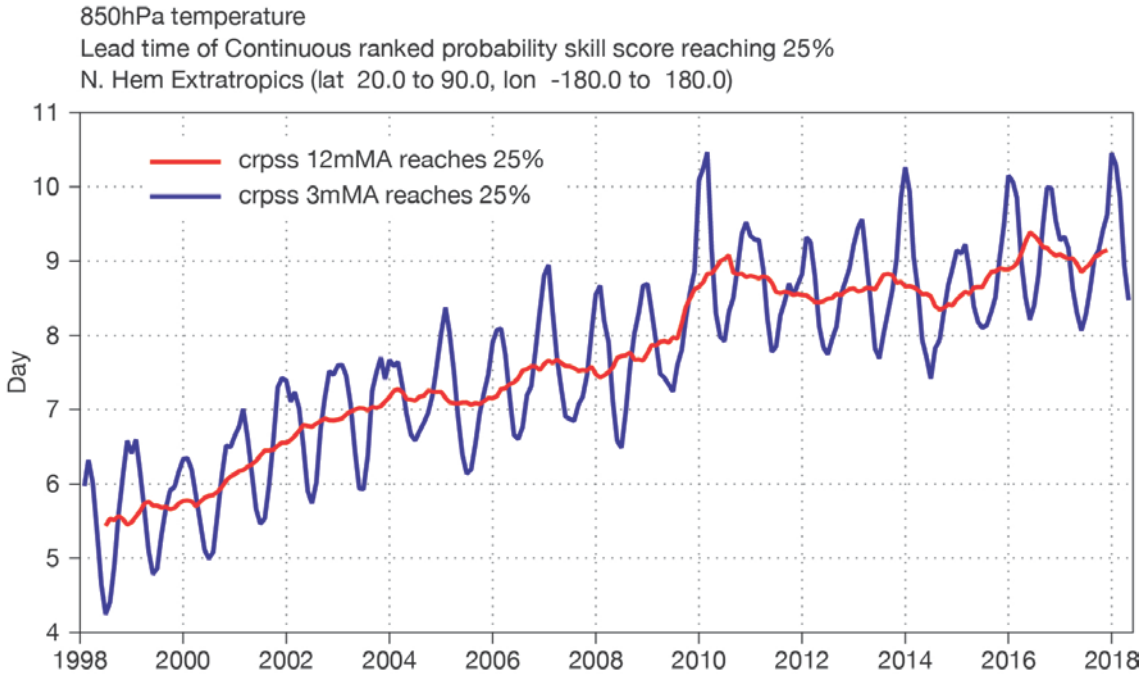

Figure 4: Headline score time series of the forecast lead time when the monthly-mean (blue line) or the 12-month-running-mean (red line) continuous rank probability skill score (CRPSS) of the ECMWF ensemble (ENS) operational forecasts of 850 hPa temperature anomalies over the Northern Hemisphere (NH) crosses the 25% threshold, indicating a gain of about 2 days per decade of ENS probabilistic predictive skill.

Figure 5: Supplementary headline score 12-month averaged timeseries (red line) of the forecast lead time at which the SEEPS score for HRES forecasts of 24-hour total precipitation reaches 45% for the extra-tropics (Northern and Southern hemispheres); verification against station observations. Also shown are the equivalent timeseries for the two reanalyses, ERA-Interim and ERA5, highlighting the component of the score that is due to interannual variability in the meteorology.

In addition, scores that are sensitive to the activity (variability), such as the root-mean-square-error (RMSE), may degrade for an under-active variable that is improved towards the observed variability. It is important that the activity of the model throughout the forecast range is close to the analysis. Only by improving the fidelity and realism of all individual components of the forecast system and their interactions, will the resulting forecasts continue to improve over the long-term. Assessment of the evolution of errors over multiple cycle implementations is therefore an important component of the evaluation, to monitor the long-term performance and feed any information on degradations back into the development process.

Figure 4 shows an example of a long timeseries of the 850 hPa temperature headline skill score for the ENS with an annual cycle and year to year variations due to meteorological variability but a long term increase in skill. The interannual meteorological variability can be taken into account by using the reanalysis as the reference. Figure 5 shows an example for the supplementary headline score of HRES precipitation evolving over the last 15 years. The equivalent timeseries for both the ERA-Interim and ERA5 reanalyses are also shown to highlight the component that is due to variations in meteorological regimes.

3.2. Practical implementation of the IFS evaluation

This section provides information on the practical application of many of the principles discussed in Section 3.1 to the evaluation of forecasts for the medium-range, extended-range and seasonal timescales.

The efficiency of the testing and evaluation process is vital, and increasing automation and ease of use of evaluation software is an ongoing part of the continuous improvement process. Initial quality control

to catch bugs and unphysical behaviour is the first step, and can be performed with computationally cheap, low-resolution configurations. Recently introduced fast-turnaround automated tests within the code management environment also help in this respect, saving developer's time and resources. The scientific evaluation of the forecasts then follows with some standard verification for all configurations but with emphasis on different characteristics depending on the forecast lead-time.

3.2.1. Software tools

There are several software tools that are used for operational verification, new IFS cycle evaluation, and during the IFS development process, with statistics, plots and summary scorecards available on the internal web. Although developed in-house with different histories, many are used across the Research and the Forecast Departments (RD and FD).

The main verification software packages in regular use are:

- Verify/Quaver – Main operational verification system for the medium and monthly ranges for atmosphere and waves, HRES and ENS. Also used in RD.
- EPSverify – ENS verification partly integrated in Verify, being extended for multiple-resolution ensembles, fair scores and maps of probabilistic score differences.
- Obstat – Observation based statistics for analysis and forecast departures (using ODB software).
- IVER – Widely used in RD for the analysis and deterministic medium-range forecasts, automatically creating user-friendly web page with hundreds of plots.
- Ver0D – Verification against static point-wise observations and time series of station data (used in CAMS).
- The Diagnostics Toolbox – Flexible set of tools that can be applied in model space and observation space to produce diagnostics related to forecast error, and diagnostics based on important budgets associated with mean initial tendencies and analysis increments, and EDA variances. The Diagnostics Explorer is a series of web pages to visualise the results.
- Bespoke software for monthly and seasonal forecast evaluation.

There is some overlap between the functionalities of the different packages, but each has some unique characteristics, and given the considerable development effort for each, replacing all functionality with a single unified comprehensive software package is not considered a worthwhile use of resources. However, an activity began in 2017 (the Verification Project) to look at the current verification infrastructure, data structures and workflows. The initial focus is on some level of convergence starting with software components that can be shared to reduce duplication and maintenance. While the evaluation software for the monthly forecasts is functional, the software for evaluation of seasonal is less mature and needs further work.

3.2.2. Medium-range forecast evaluation

For the medium-range, the evaluation of forecast skill involves a comprehensive set of verification metrics applied to the HRES and the ENS for both upper-air and surface fields, and against own analysis and against observations. The primary metrics for single, deterministic simulations (at the operational,

high-resolution or at lower-resolution) are the RMSE and mean error (bias), anomaly correlation (ACC), and contingency-table based scores (SEEPS, equitable threat score, frequency bias) for surface fields such as precipitation. For the ENS, the primary metrics are the CRPS/CRPSS and the mean error-spread skill. Brier scores and skill scores, and the area under a Relative Operating Characteristics curve (ROCA) for the probabilistic prediction of categorical events are also used. The reader is referred to Wilks (2011) for a comprehensive description of these scores.

Score differences between the new and the operational cycle are evaluated, and significance tests applied (bootstrapping and/or Student's t-test). Diagnostics include maps of scores, latitude-height cross-sections of longitudinally averaged scores, and area-aggregated scores. The most high-level summary of cycle differences is provided by scorecards, which are routinely produced both for HRES and ENS (see two different examples of the ENS scorecards in Figures 6 and 7).

While the above methodology evaluates the mean changes in skill, it does not quantify the effect of a new cycle on the prediction of extremes, such as tropical cyclones. The verification of tropical cyclone forecasts is part of the evaluation but suffers from the relatively small sample provided by a typical e-suite period. The impact of changes on extremes are evaluated also looking at the reforecasts, to increase the number of cases. It should be noted that, when assessing extremes, it is important not to focus only on cases when the extreme was observed, but assess all cases in order to retain propriety of the scores and avoid introducing biases in the statistics. For strong wind and heavy precipitation in general, a more systematic evaluation of score differences could be adopted by looking at scores for higher quantiles. However, a compromise must be made between the degree of 'extremeness' and the corresponding sample size and resulting statistical significance. For these extremes, to have a sounder evaluation we complement single-case studies with statistical evaluations based on ‘less extreme but similar cases’ (e.g. less intense wind storms, or precipitation events). By contrasting the two, we can have an ‘as-good-as-feasible’ evaluation.

3.2.3. Extended range forecast evaluation

Monthly re-forecast experiments are performed to assess the impact of the physics contributions to the extended range forecasts. They consist of 15-member ensembles integrated for 46 days and starting on 1 February, May, August and November (to sample all seasons) and covering 27 years, from 1989 to 2016. The verification is performed using an extensive diagnostics package which provides an assessment of the model biases, forecast probabilistic skill scores (e.g. CRPS, ROCA) and ensemble spread for more than 20 variables, including sea-surface temperature (SST), sea-ice and tropical cyclone activity diagnostics.

As for the case of the HRES and ENS, the forecast skill scores are summarized in a scorecard, like the one produced for medium-range forecasts but for weekly mean anomalies instead (Figure 8). This diagnostic package also assesses the impact of model changes on the main sources of extended-range predictability such as the Madden Julian Oscillation, tropical-extratropical teleconnections, and sudden stratospheric warmings. Probabilistic skill scores are also applied to low-frequency variability patterns, such as the North Atlantic Oscillation (NAO) and the Pacific North-Atlantic (PNA) patters. As for the case of HRES and ENS, the time-evolution of forecast scores for the monthly timescale are monitored, to assess whether Cycle upgrades bring overall improvements (Figure 9).

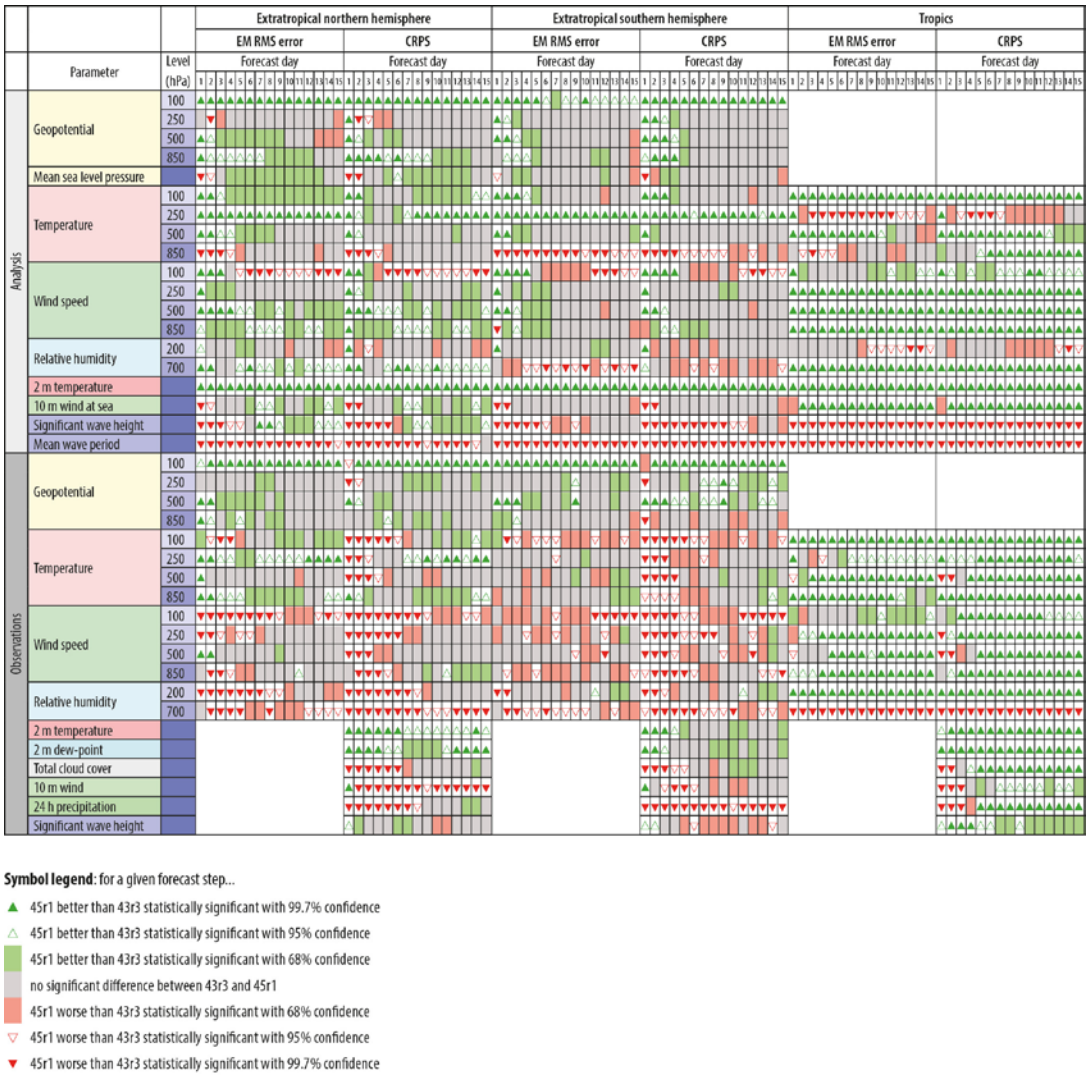

Figure 6: The scorecard used to summarize the difference between the performance of the medium-range/monthly ensemble (ENS) up to forecast day 15, based on model cycle 45r1 (the 'new' cycle that was under-testing and was implemented in June 2018), and the operational cycle 43r3. It shows, for a range of variables, levels and areas, whether the new cycle is statistically significantly better than the operational cycle.

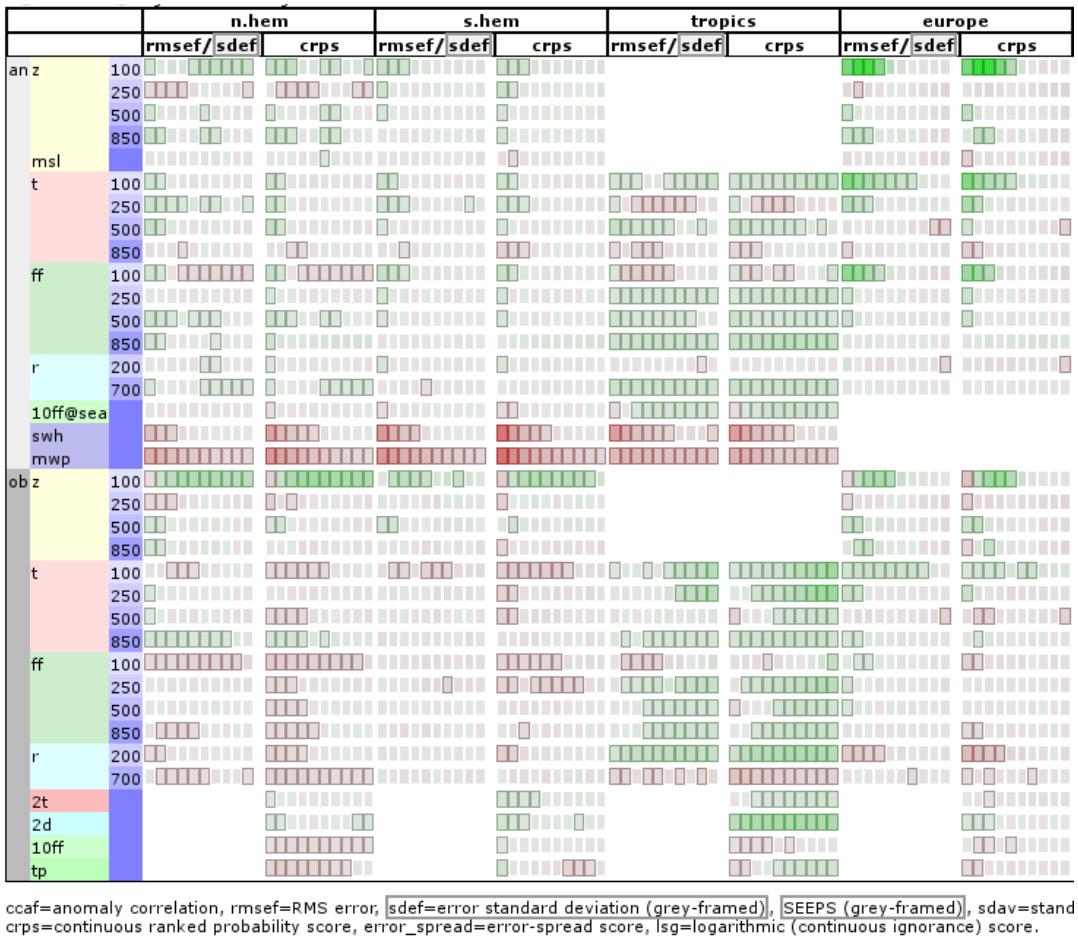

Figure 7: Similar to Fig. 6 but for the summer season difference between the ENS up to day 10 for model cycle 45r1 compared to the operational cycle 43r3. The depth of shading of the squares relates to the magnitude of the change and the boxes are highlighted if they are statistically significant.

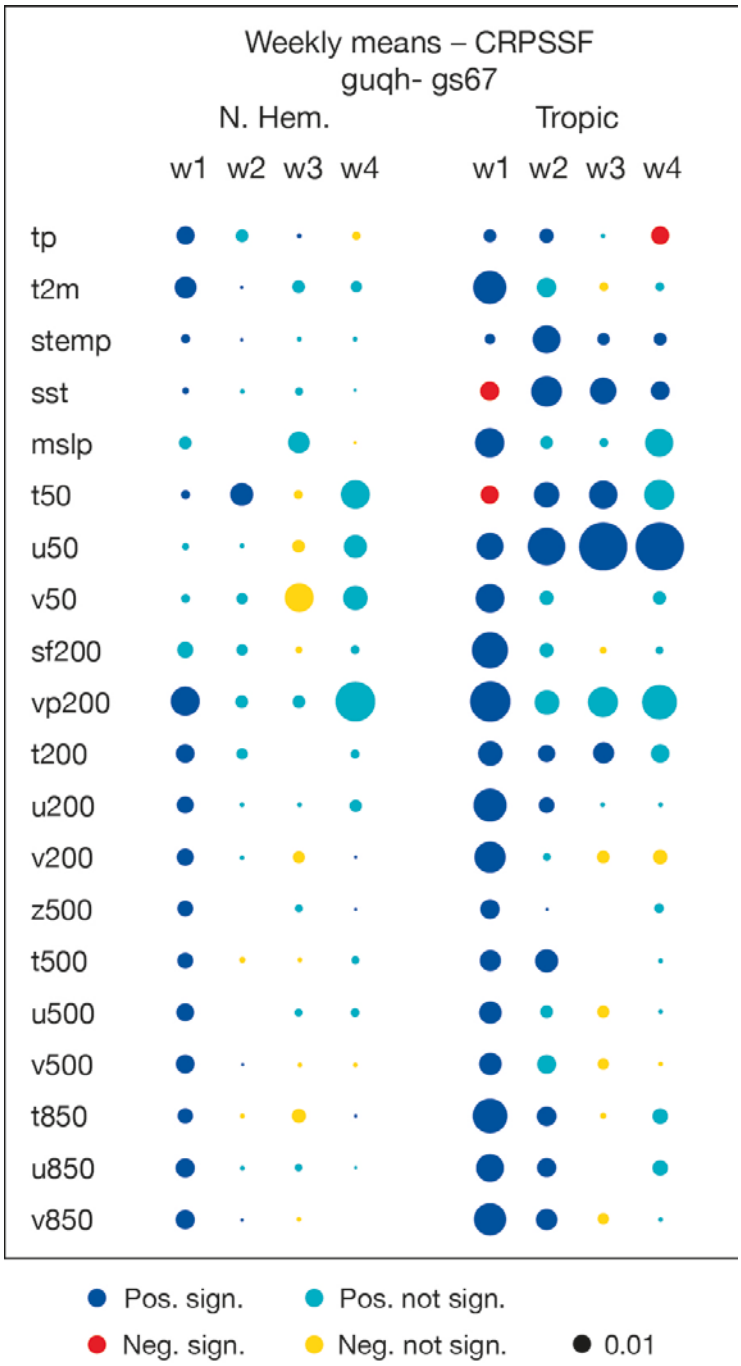

Figure 8: Score-card for the monthly forecasts generated by ENS, for IFS cycle 45r1 computed with respect to operations (cycle 43r3).

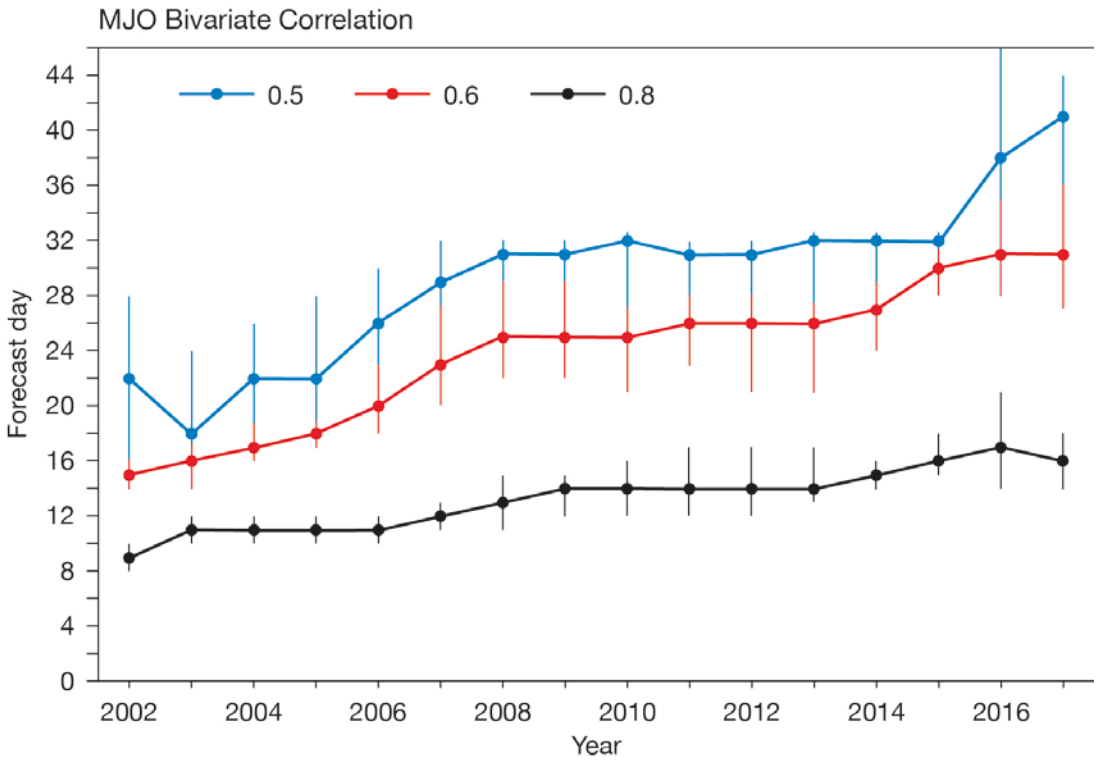

Figure 9: Time evolution of the skill of operational ENS monthly predictions of the MJO, issued by the cycles operational from 2002 to 2017. The three curves show the forecast day when the ACC of the forecast crosses the 0.5 (blue line), 0.6 (red line) and the 0.8 (brown) threshold. The reader is referred to Vitart et al (2016) for more information.

3.2.4. Seasonal forecasts and the model climate evaluation

Evaluating the climate of the model on seasonal timescales, in terms of both mean errors and variability, is an important part of the development of a new model cycle. The model always tends towards its own climate, so continuing to reduce systematic errors and capture the observed distribution and variability across a range of quantities is important, not only for the seasonal timescale, but also for the medium-range and monthly forecast times.

The model climate is first evaluated for each forecast model change in a computationally-efficient low-resolution (TL255L137, 80km grid resolution), coupled 4-member ensemble of 1-year free-running forecasts ("climplot"). Seasonal and annual means are evaluated against the reanalysis and a range of satellite datasets (e.g. radiation, cloud, precipitation, water vapour, wind).

Combined changes are then evaluated by applying a more comprehensive package on seasonal forecasts, based on a small ensemble of 7-month coupled forecasts initialised the 1st of May and November, for 20 years, from 1981-2010, allowing all four seasons to be evaluated. All IFS cycles from 37r3 and onward have been evaluated with a fixed resolution, which is TL255L91 for the atmosphere and the ORCA1 (1-degree resolution, 42 vertical levels) resolution for the NEMO ocean model. For an example of the evolution of the model climate for these cycles, the reader is referred to Stockdale et al (2018 - SAC paper on seasonal forecasting). In the above testing suites, the stochastic physics is switched off.

The model climate is evaluated using an extensive diagnostic package using ERA-Interim (to be upgraded to ERA5 as soon as it completes the period from 1981 to the current date) and independent data sets where available. This package produces many scores, and around 2000 plots. Among the mean aspects that are assessed are upper-air fields such as temperature and winds, sea-surface temperature, radiation and precipitation. The variability is evaluated with scale-dependent standard deviations together with phenomena like the MJO (Madden-Julian Oscillation), ENSO (El Nino Southern Oscillation), teleconnections, blocking and flow regimes. The plots from the final model climate evaluation are available on the ECMWF intranet (in the Diagnostic Explorer) and the results are also summarized in a scorecard, which is less mature and expected to evolve in the future.

3.3. Using the evaluation to further develop the IFS

For the continuous improvement of the IFS, it is important to have strong links between the ongoing evaluation of the operational configurations, evaluation of each new IFS cycle, longer-term diagnostic work, and the IFS research and development process and plans.

Identifying, documenting and communicating problem areas help to inform and prioritize the research and development needed to address these aspects in future cycles. This includes information of past and current performance of the IFS on the web, cross-department projects targeting specific problem areas and feedback from users of the operational forecasts.

3.3.1. A record of performance

Today's technology allows the creation of vast amounts of information, but also the potential to lose easily that information over time. It is important for the continuing development of the IFS to keep a sufficient record of the evaluation of each cycle as well as more in-depth diagnostics of changes to the IFS. There is valuable information from past changes that can help to inform future developments and avoid duplication of results. Currently this takes many forms, including a summary of the changes in each cycle on the external web, both descriptively and as a summary scorecard with external users as the target audience (www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model).

Changes are also summarized in Newsletter articles, which target the ECMWF users' community, and some of the more significant scientific and technical developments are described in Technical Memoranda or submitted to international journals. Plots charting the evolution of the model climate over cycles dating back to 2003 are available on the web.

The Diagnostics Explorer is an internal set of web pages providing easy access to many diagnostics on the analysis and forecasts over a number of recent cycles. More in-depth information on the changes to each cycle are stored on the internal web in the form of an automatically generated summary document (the "FLUB"). Since cycle 43r3, all changes to the IFS from individual developers are stored as tickets in issue-tracking software (JIRA) with details on the changes to the code, testing and impacts.

So, there are many forms of record of the changes to the IFS software and performance. There is potential for improvement as some diagnostics, such as scores for thematic merged contributions, that would be a useful reference when looking back at the impacts of changes to previous cycles, are not always kept in a systematic way. To help to address this, an RD Memo is now created for each Cycle describing the main contributions and their impacts in more detail as a reference for the future.

3.3.2. Topical/targeted projects

Topical, targeted cross-department projects aimed at investigating specific aspects are sometimes initiated when there is enough evidence of a need for a more thorough investigation.

These projects take different forms depending on the problem being addressed, but the purpose is to target a specific aspect of the forecast by bringing together the relevant experts within ECMWF, and sometimes outside, to make a step-change in understanding and improvement. Evaluation of the IFS forecasts from different operational configurations and for IFS cycle upgrades has provided the motivation and information for these activities. Two examples that are currently ongoing are 'USURF' (Understanding uncertainties in surface-atmosphere exchange) to better understand the causes of biases in near-surface weather parameters in the IFS, and the 'Stratospheric Task Force' to address longstanding systematic errors in the middle atmosphere.

There is a clear benefit of such targeted projects in bringing experts from different areas together to work on a common problem, potentially leading to new insights and eventual improvements to the forecast system. It is important to keep the group focused. Further topical projects will be established in the future as appropriate.

3.3.3. Operational forecast feedback

An important source of information on problem areas in the operational forecasts is feedback from users in National Meteorological Services and from commercial customers. This is either reported directly when an issue is found (via forecast_user@ecmwf.int, Service Desk or Data Services), or annually through written reports or at user/forecaster meetings at ECMWF (e.g. the "Using ECMWF's Forecasts" meeting, held annually in June at ECMWF). There is a web-based User Request Management System for recording and processing external user requests and these can inform the IFS development.

An ECMWF analyst is on duty every weekday to assess the daily forecasts, and perform an initial investigation of any problems they identify or that are reported by external users. The analyst provides a summary of these investigations in the Daily Report, made available internally to ECMWF staff at the end of each day. Already at this stage the assessment often involves people from the Research Department. The main issues that point towards deficiencies in the forecasts, rather than random errors, are explained at weekly weather discussions open to all staff.

Quarterly Evaluation and Development (QED) meetings review the performance over the previous season and all staff are encouraged to attend. Actions from QED meetings are recorded and followed-up if there is a possible solution. The issues often result in research activities with solutions making it into new IFS cycles. Unresolved issues are included on the publicly available "Known IFS forecasting issues database" to inform users about forecasts deficiencies and feed into longer term ECMWF research plans.

4. Future evolution of the evaluation process

In the previous sections, it has been shown that at all stages of testing there is a trade-off between cost and realism of the simulated impact, and during the process going from initial component tests to final

e-suite these tests progressively form a closer approximation of all aspects of the final operational configurations.

As the IFS evolves, the evaluation process also needs to evolve to test all configurations, whilst remaining affordable and manageable. In this section, we discuss the next steps in this evolution. Areas where costs of fully replicating operations are high, yet additional early testing is beneficial include:

- Ensemble of Data Assimilation (EDA) changes and feedback;
- Ensemble (ENS) forecast for medium and extended range;
- Seasonal (SEAS);
- Coupled model;
- Coupled data assimilation;
- Impact of changes in a reanalysis context.

In this section, we consider what additional evaluation could be done in all areas, especially at early stages. An overarching requirement is that we do not wish to impose too prescriptive a testing procedure, that will force many unnecessary tests to be done. The ideal is to have a system that can run easily a minimum baseline set of tests, and then have a toolbox of additional tests, which can be run at later stages of testing, or if earlier testing raises concerns. The judgment of individual scientists about what is needed will remain a critical part of the process.

4.1. Discussion of testing options

In this section, we discuss changes for each component and present a flow diagram (Figure 10) for the testing process. Tables providing a complete description of proposed testing in each stage, including that already done, are included in Appendix B.

4.1.1. VarBC spin-up issues in HRES

Single runs designed to evaluate impact in the HRES configuration are well established, but there remain difficult aspects, such as slowly evolving error components of the atmospheric analysis, the variational bias correction coefficients, and the model error. These components are resolution-dependent and have long spin-up times (months). The solution adopted to mitigate these problems has been to initialise test experiments starting from already spun-up experiments at same or similar resolution, so that approximate stationarity in error evolution can be expected. However, these aspects are still not fully understood, and are a topic of current research, and therefore the best approach remains an open issue.

4.1.2. Slim EDA configuration for testing

To gain early insight into the likelihood of interactions of changes in the model and/or the observation usage with the EDA, new cheaper EDA testing configurations have been developed: a 10-member TCo399 RD test configuration and a 10-member TCo639 configuration aimed at testing impacts during the merge stage. This “slim EDA” is described in more detail in the Advancing Weather Science report. It provides results qualitatively coherent with the full resolution system at a cost affordable in early stage testing

Diagnostics of the "slim EDA" will be evaluated to see whether there is strong sensitivity, e.g., in terms of changes in the spread variations, correlation scales, etc. Once these indicate strong sensitivity, further diagnostics, evaluation, or testing, also in the context of deterministic 4D-Var, might be needed. A full EDA will continue to be run in the final stage of pre-operational tests.

4.1.3. Reduced Ensemble configuration for testing

Changes to the skill of the ensemble forecasts can be due to many factors. To better trace the cause of impacts and improve the feedback in the development process, a more detailed testing approach is recommended for the medium-range ensemble. This could be done with modest additional computational resources by using an intermediate horizontal resolution (e.g. TCo399) and a reduced ensemble size (e.g. 8 members). As demonstrated by Leutbecher (2018), changes in probabilistic skill for the full ensemble size of 50 can be estimated from small ensembles using score adjustments. This can then allow earlier evaluation of changes on the ENS system.

4.1.4. Full seasonal skill evaluation of every model cycle

To test more effectively each new cycle on the seasonal time scale (i.e. using the seasonal ensemble SEAS) two important aspects need revisiting: i) updating the low-resolution configuration currently used; and ii) conducting a full seasonal evaluation for each model cycle, rather than every few years (aligned with the SEAS production cycle).

It is proposed to update the seasonal forecast testing-suite as follows: the ocean will use the ORCA1_Z75 configuration, which has the same vertical discretization as the operational ocean but lower horizontal resolution. The low-resolution ocean will be initialized by the 0.25-degree ORAS5, and the atmosphere will be initialized by ERA5. The atmospheric resolution will be TCo199L91 and it will include stochastic physics. The reforecast period will cover the same period used to generate the SEAS5 re-forecasts, 1981-2016. This configuration will be used to evaluate model fidelity, and can also be used to test some elements of forecast skill related with SST and sea ice. In addition, a high-resolution run is needed for cycles proposed for new SEAS versions. In both cases long testing periods are needed because of the timescales involved.

4.1.5. Coupled forecast with 1-degree ocean

From cycle 46r1, testing the coupled atmosphere-ocean has been included as a requirement in RD tests, but running the ocean at operational resolution (0.25 degree) is too costly. As mentioned above, a methodology has been developed to initialize the low-resolution ocean model from a high-resolution ocean reanalysis: ORAS5. This approach needs further evaluation to confirm its suitability for routine testing, but is a promising development to provide an option for coupled ocean-atmosphere forecast tests that is less expensive.

4.1.6. Coupled Assimilation

As noted above, since cycle 45r1 we have run a coupled ocean-atmosphere forecast model in test experiments, and methods for coupled assimilation testing are therefore also being developed. Next year, starting with cycle 46r1 we will evaluate the relevance of running the deterministic and EDA test experiments in weakly coupled assimilation mode. In the longer term, the testing will be extended to

include a Quasi Strongly Coupled Data Assimilation (QSCDA) testing environment evaluation. Testing this configuration will pose additional challenges because a) error dynamics in the ocean evolve on longer time scales with respect to the atmospheric dynamics and the corresponding data assimilation window, and b) running the coupled ocean data assimilation component at reduced resolution risks losing the main sources of spatial variability of the errors. These problems are like the issues encountered initializing the slowly evolving error components of the atmospheric analysis. It remains an open issue how testing of this should be undertaken, because feedback could occur over very long time-periods.

The testing of new ocean model or data assimilation components is not covered here. Suffice it to say that time-scales needed to test the ocean are multi-year or multi-decade, and they need long ocean-only or coupled integrations.

4.1.7. Testing for Reanalysis

It is worth a few words about the reanalysis suite, which now (i.e. in the current ERA5 reanalysis under production) includes a single, high-resolution analysis and an ensemble of data assimilation component. Every 7-10 years, a new reanalysis is generated, so that all model and data-assimilation upgrades, and a high-resolution, can be used to provide a better estimate of the past. Thus, we need to make sure that, when the times come to start production of a new reanalysis, the existing (possibly the operational) model cycle can be used.

Reanalysis needs to provide optimal results for several decades in the past, where the observing system is much sparser, and for which the optimization of IFS towards the current observing system may be sub-optimal. It is, however, impractical to assess this for each new model cycle.

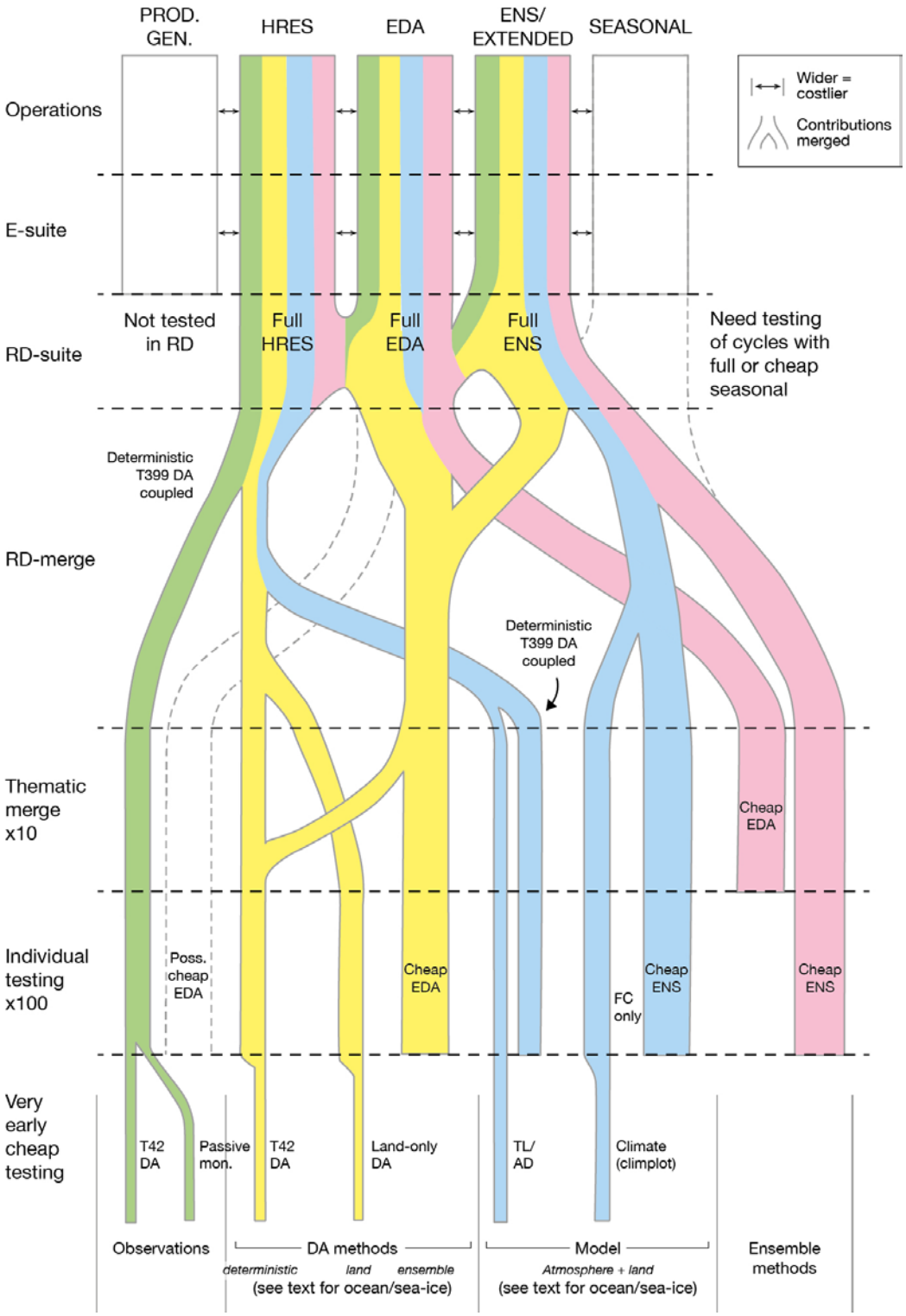

Figure 10: Schematic of the ECMWF testing hierarchical strategy followed to build new cycles of the ECMWF Integrated Forecasting System (IFS): width highlights the cost and the colour denotes the IFS component. Testing starts (at the bottom) with 'Very early cheap testing' of the individual components, and ends (at the top) with the 'E-suite' and 'Operations' testing performed at full resolution, including all inter-dependencies.

What should be tested is the technical ability to run the system for past periods (i.e., ensure that required infra-structures are not broken) at an early stage after the delivery of a model cycle. This should be an inter-departmental activity, performed mainly between RD and the Copernicus Department (COPD). Given that reanalysis production is now the responsibility of COPD, it should be the responsibility of the C3S Reanalysis Team to test the technical ability of each model cycle. In addition, as a stress test, RD developers should be encouraged to also test their developments using a reduced part of current-day data availability. This is an area of ongoing discussion, aimed at developing the most effective testing strategy.

4.1.8. Testing hierarchy

The testing of a new cycle includes a hierarchy of tests, as described in section 2. A schematic of the flow and inter-relationship of these tests as applied to the main IFS components is shown in Figure 10. The actual testing configuration of the main stages highlighted in Figure 10 are listed in details in the Tables 5-8 reported in Appendix B. The main stages, as discussed earlier, include:

- Individual component tests (Alpha-1a testing; see Table 5 for details);
- Thematic merge tests (Alpha-1b testing; see Table 6 for details);
- Research Department merge tests (Alpha-2 testing; see Table 7, for details);
- RD e-suite (Alpha-3 testing; see Table 8 for details).

Decisions on what should progress to the next stage of testing are typically taken by the relevant RD Section Heads in the early stages of testing, and by the full RD Management Team in the latter stages (with consultation with FD as required).

In addition, it should be noted that some components are only tested in FD operational configuration (e.g. satellite simulated images, product generation), i.e. after all RD testing has been completed.

The final decision on whether the new cycle is ready for operational implementation is taken by FD, after the RD and FD e-suites have covered a large enough sample of cases (typically at least 2 seasons of analyses and forecasts) and have been comprehensively evaluated.

4.2. Directions for future development of the evaluation process

The evaluation process for IFS cycles has evolved significantly over time, with increasing complexity of the system and increasing scrutiny of a growing range of output from the IFS. Recent cycles have been more comprehensively assessed than ever before. However, there are always improvements to be made, particularly in the context of extending verification to wider aspects of the Earth-system (atmosphere, land, ocean, sea ice, rivers, lakes, atmospheric composition), adapting techniques for multi-resolution ensembles, new metrics and methodologies and extracting more from the observations that are available. This section follows on from the evaluation discussion in section 3 and identifies a few specific areas of the IFS evaluation process being considered for the future and where further work might be needed to inform the future directions.

4.2.1. A more comprehensive evaluation against observations

Maximizing the use of information in the observations is a key area for improvement. Assessing forecast quality in observation space is the best way to complement the verification in model space, against analysis. The plan is:

- To extend the verification of the analysis against observations, to the forecast. Ways to make this easier and computationally cheaper will be explored and whether this framework could also be applied to the ensemble (EDA, ENS).
- To extend the range of observations in the ODB (Observation Database) for monitoring, in particular satellite observations, and including non-real-time observations, to make automating the verification easier.
- To account for observation error in the verification statistics.

4.2.2. Improving flow-regime dependent diagnostics

Evaluation techniques based on flow-regimes help to target the identification of regime-dependent systematic errors. New developments planned for the Diagnostics Toolbox will allow the user to objectively calculate a set of clusters based on initial flow-types prior to applying the tools. This should help the user identify and prioritise future research efforts for model improvement.

4.2.3. Improving diagnostics of high-impact weather

There is an increasing number of model products aimed at forecast users, some related to high impact weather (e.g. wind gusts, pseudo-satellite imagery, precipitation type, visibility, lightning). Although they are diagnostic and do not affect the model evolution, it is clearly important that the quality of these products is part of the evaluation. There have been occasions when errors have slipped through and been spotted only very late in the process or even after operational implementation. More comprehensive and routine evaluation of these types of products would ensure the ongoing quality for each new Cycle release.

4.2.4. Efficiency of the evaluation software

The current ongoing Verification Project aims for better coordination and cooperation on verification across ECMWF. The initial focus is on the data management aspects of verification that are common to all software packages. There is planned work to improve the verification workflow and improve the efficiency of the development, maintenance and use of software packages. The overall aim is to reduce duplication and use common components of the software where it makes sense to do so, whether for data access, calculations of scores or use of observational data.

4.2.5. Updating and extending metrics

With increasing emphasis on the ensemble, additional evaluation metrics may need to be considered. The traditional RMSE for single forecasts, and CRPS for ENS are both sensitive to the amplitude of the error and other metrics in the standard verification packages could be explored to help inform the development process. One example is for the evaluation of precipitation. Given the large uncertainties in verifying precipitation, a categorical skill score (SEEPS) with climatologically dependent dry, light

and heavy precipitation categories, was devised as a supplementary headline score for HRES. An equivalent skill score could be applied to ENS. Several verification statistics currently use ERA-Interim as a reference and these will be updated to use ERA5 when fully available.

4.2.6. Extending evaluation methods to wider aspects of the Earth-system

All IFS configurations are now coupled with the ocean and sea-ice models, and yet the evaluation of the ocean and sea-ice is relatively limited at present. This is in part due to fewer available observations compared to the relatively well-observed atmosphere, but also the lack of software infrastructure. The initial aim is to improve the verification of SST and sea-ice against available analyses and observations in the standard verification packages. Including the ocean analysis, ORAS5, and/or ocean observations in the ODB will facilitate this.

For the longer-range forecasts, there is a need to evaluate relationships between the mean state and variability and predictive skill, e.g. stratosphere-troposphere interactions, polar-midlatitude connections, the influence of SST and land surface. Further statistics of phenomena are also needed, such as frequency of regimes or position of storm tracks, like the current statistics of tropical cyclones or MJO. This is an ongoing research topic with the potential for external collaboration, but is needed to help to identify future improvements for extended-range predictions.

5. Conclusions

This paper has reviewed the testing strategy for bringing together multiple candidate developments to produce a new cycle.

The complexity of the model cycle development and evaluation process has been increasing substantially in the past few years (e.g. with the introduction of the EDA, and its use in 4D-Var and ENS; and with the coupling to the NEMO 3-dimensional ocean and LIM2 sea-ice models), and regular review of the process has led to the introduction of clearer development and testing strategies.

In summary:

- ***The IFS cycle development and evaluation process***, includes three main testing phases:
  1. *The alpha-phase testing*, which includes five testing stages:
     - 0 - Individual *ad-hoc* testing
     - 1a - Individual testing against controls
     - 1b – Thematic pre-merging and addressing interactions
     - 2 - Incremental build and testing
     - 3 - RD e-suite: higher resolution HRES/ENS and EDA testing
  2. *The beta-phase testing*;
  3. *The release-candidate-phase testing;*
- ***Evaluation involves a range of methods and metrics***, applied to analyses and forecasts, with scores summarized, when meaningful, in scorecards;

- ***Testing is hierarchical***, organized in stages of increasing complexity.

Inevitably the testing process always has been - and will need to continue to be - a hierarchical one. However, a number of considerations arising from the strategic drivers and increasing complexity of the system have led us to propose (and to some extent already implement) some evolution in the tests and evaluation performed in the alpha-phase. Three areas in which work is focussing to upgrade further the testing strategy are summarized below:

- Increased focus on ensembles;
- Increased focus on the extended range;
- Developing cost-effective methods to test the increasingly complex IFS

Increased focus on ensembles. With the strategic focus on ensemble predictions, it has been recognized that the balance between HRES and ENS testing was sub-optimal (with routine ENS testing historically being typically carried out only relatively late in the development process). This issue has already been addressed in the development of cycle 46r1 with the addition of extra ENS tests at an early stage (Section 2). However, that first implementation of enhanced more routine ENS testing, while a positive step, almost certainly did not strike the optimum configuration for cost-effective testing. Accordingly, a further iteration is planned, seeking to retain the increased focus on ENS performance, while optimizing the detailed approach (e.g. by testing with a smaller ensemble and taking advantage of fair scores to predict impact of a change on the full ensemble). More generally, in evaluating the tests, we will ensure that an appropriately broad range of measures is considered and weighted appropriately (e.g. increased emphasis to near-surface 'weather').

Increased focus on the extended range. The desire to improve monthly and seasonal predictions (and to take a seamless approach for the benefit of all timescales) has led us to question whether the testing strategy for the more extended predictions is adequate. For the monthly system, moves in recent years have already made early testing of model changes on this timescale much more routine, and we judge that the priority is to continue with this approach, while giving slightly more emphasis to these results in the overall holistic assessment of a cycle. On seasonal timescales, full testing is extremely expensive, but we have proposed a new hierarchy of tests that should be employed to provide extra insights. Specifically, these include a low-resolution set-up that can realistically be used to give an assessment of model climate and ENSO performance early in the development process, and a full resolution configuration (but with a limited set of start dates) for fuller evaluation of a cycle.

Developing cost-effective methods to test the increasingly complex IFS. As noted in the introduction, the full operational system has been becoming increasingly complex in recent years e.g. through coupling between the EDA and the 4D-Var, and through all forecast systems now being coupled to an interactive ocean (and sea-ice). This has made it appropriate to review whether our testing strategies remain appropriate. On the former issue, we have concluded that unsatisfactory risks (of late surprises) were being introduced through the fact that testing the full EDA is only affordable very late in the development process. Accordingly, a new 'slim EDA' has been developed to allow earlier sight of likely signals, and work in the coming year will seek to embed (and refine) its use as part of our testing strategy. On the ocean side, 'workhorse' testing at 46r1 for all timescales has used a coupled system (which seems appropriate), but the current setup uses a 0.25-degree ocean (as operations) while the atmosphere

resolution is typically degraded from that used in operations. This feels out of balance, and work in the coming year will seek to confirm that tests with a 1 degree ocean will be a good enough predictor of the impacts of a change, that this resolution can be adopted for most early tests (across timescales). It will also be important to continue efforts to ensure convergence of evaluation software over time.

We believe that these changes represent a natural evolution of our testing strategy rather than a revolution. Through making them - and being open to further emphasis shifts as we gain experience of working with them - we are confident that we can continue to deliver the developments required to keep the ECMWF systems world-leading.

6. References

Bernard B., G. Madec, T. Penduff, J.-M. Molines, A.-M. Treguier, J. Le Sommer, A. Beckmann, A. Biastoch, C. Böning, J. Dengg, C. Derval, E. Durand, S. Gulev, E. Remy, C. Talandier, S. Theetten, M. Maltrud, J. McClean and B. De Cuevas, 2006: Impact of partial steps and momentum advection schemes in a global ocean circulation model at eddy-permitting resolution. Ocean Dynamics, 56(5-6):543–567 (ISSN 1616-7341; doi: 10.1007/s10236-006-0082-1).

Bonavita, M., Y. Trémolet, E. Hólm, S.T.K. Lang, M. Chrust, M. Janiskova, P. Lopez, P. Laloyaux, P. de Rosnay, M. Fisher, M. Hamrud and S. English, 2017: A Strategy for Data Assimilation. ECMWF Research Department Technical Memorandum n. 800, pp. 42 (available from ECMWF, Shinfield Park, Reading RG2-9AX, UK).

Bouillon, S., M.A. Morales Maqueda, V. Legat and T. Fichefet, 2009: An elastic-viscous-plastic sea ice model formulated on Arakawa B and C grids. Ocean Modelling, 27, 174-184 (doi : 10.1016/j.ocemod.2009.01.004).

Buizza R. and T. N. Palmer, 1995: The singular vector structure of the atmospheric general circulation. J. Atmos. Sci., 52, 1434-1456.

Buizza, R., E. Andersson, R. Forbes and M. Sleigh, 2017: The ECMWF Research to Operations (R2O) process. ECMWF Research Department Technical Memorandum n. 806, pp. 16 (available from ECMWF, Shinfield Park, Reading RG2-9AX, UK).

Fichefet, T. and M.A. Morales Maqueda, 1997: Sensitivity of a global sea ice model to the treatment of ice thermodynamics and dynamics. Journal of Geophysical Research, 102, 12,609-12,646, doi:10.1029/97JC00480.

Geer, A., 2015: Significance of changes in medium-range forecast scores. ECMWF Research Department Technical Memorandum n. 766, pp. 31 (available from ECMWF, Shinfield Park, Reading RG2-9AX, UK).

Isaksen, L., M. Bonavita, R. Buizza, M. Fisher, J. Haseler, M. Leutbecher and L. Raynaud, 2011: Ensemble of data assimilations at ECMWF. ECMWF Research Department Technical Memorandum n. 636, 48 pp (available from ECMWF, Shinfield Park, Reading RG2 9AX).

Janssen, P.A.E.M., O. Breivik, K. Mogensen, F. Vitart, M. Balmaseda, J.-R. Bidlot, S. Keeley, M. Leutbecher, L. Magnusson and F. Molteni, 2013: Air-sea interaction and surface waves. ECMWF Research Department Technical Memorandum n. 712. Available from ECMWF, Shinfield Park, Reading RG2-9AX (see also http://old.ecmwf.int/publications/).

Leutbecher, M., 2018: Ensemble size: how suboptimal is less than infinity? Q. Jou. Roy. Meteorol. Soc., in press (doi: 10.1002/qj.3387).

Madec, G., 2008: Nemo ocean engine. Note du pôle de modúelisation. Institut Pierre-Simon Laplace (IPSL), France, no 27, ISSN No. 1288-1619.

Mogensen, K., M. Alonso Balmaseda and A. Weaver, 2012a: The NEMOVAR ocean data assimilation system as implemented in the ECMWF ocean analysis for System 4. ECMWF Research Department Technical Memorandum n. 668, pp 59 (available from ECMWF, Shinfield Park, Reading RG2-9AX).

Mogensen, K., S. Keeley and P. Towers, 2012b: Coupling of the NEMO and IFS models in a single executable. ECMWF Research Department Technical Memorandum n. 673, pp. 23 (available from ECMWF, Shinfield Park, Reading RG2-9AX).

Molteni, F., R. Buizza, T.N. Palmer and T. Petroliagis, 1996: The new ECMWF ensemble prediction system: methodology and validation. Q. J. R. Meteorol. Soc., 122, 73-119.

Rabier, F, H. Järvinen, E., Klinker, J.-F. Mahfouf and A. Simmons, 2000: The ECMWF operational implementation of four dimensional variational assimilation. Part I: Experimental results with simplified physics. Q. J. R. Meteorol. Soc., 126, 1143–1170.

Stockdale et al, 2018: SEAS5 and the future evolution of the long-range forecast system. A paper submitted to the 47th Session of the Scientific Advisory Committee [ECMWF/SAC/47(18)11].

Vitart, F., G. Balsamo, R. Buizza, L. Ferranti, S. Keeley, L. Magnusson, F. Molteni and A. Weisheimer, 2014: Sub-seasonal predictions. ECMWF Research Department Technical Memorandum n. 738, pp. 45 (available from ECMWF, Shinfield Park, Reading RG2-9AX, UK).

Zuo H., M.A. Balmaseda, K. Mogensen and S. Tietsche, 2018: OCEAN5: the ECMWF Ocean Reanalysis System and its Real-Time analysis component. ECMWF Research Department Technical Memorandum n. 823 (available from ECMWF, Shinfield Park, Reading RG2-9AX, UK).

Wedi, N., P. Bauer, W. Deconinck, M. Diamantakis, M. Hamrud, C. Kuehnlein, S. Malardel, K. Mogensen, G. Mozdzynski and P. Smolarkiewicz, 2015: The modelling infrastructure of the Integrated Forecasting System: Recent advances and future challenges. ECMWF Research Department Technical Memorandum n. 760, pp. 48 (available from ECMWF, Shinfield Park, Reading RG2-9AX).

Wilks, D., 2011: Statistical methods for atmospheric sciences, 3rd Edition. International Geophysics Series, Vol. 100. Academic Press (Elsevier), pp. 661 (ISBN 978-0-12-385022-5).

Appendix A – Reforecast suites' testing strategy

Tables 4a-4b-4c give a summary of the ENS and SEAS5 reforecast configurations used in the evaluation of the ENS monthly forecast range, and the seasonal forecast range. All the integrations are conducted with the coupled model. The unit for the cost is a medium-range-equivalent (MRE): the cost of a single 15-day coupled forecast member. See text for details on the evaluation criteria. By comparing Tables 4b-c with Table 4a, one can see that by using reduced configurations, the reforecast cost is significantly reduced, and thus the time required to complete a reforecast evaluation decreases substantially.

Tables 4c-4b-4c highlight the key characteristics of the three main kinds of evaluation that are applied to the ECMWF reforecasts:

- Table 4a lists the Tier (A) evaluation of the skill using the whole operational re-forecast set, and can be made only after a new cycle is operational.
- Table 4b lists the Tier (B) evaluation conducted with the final configuration chosen for operational update, at full resolution but with a reduced sample in the re-forecast data set. In the case of seasonal forecasts, the intention is that this level of testing is done for every operational medium-range cycle. This enables tracking of the seasonal performance of successive IFS cycles.
- Table 4c lists the Tier (C) evaluation, an even slimmer test suite, chosen to allow testing of specific updates or proposed changes. The model resolution and ensemble size are reduced. Appropriate options for future Tier (C) seasonal testing are still being discussed, as described later in this document.

The diagnostics used have different levels of scientific maturity, and only a few of them can be condensed into objective and discerning metrics. The other aspect is the readiness of the diagnostic infrastructure. We have tried to summarise these aspects in the entry "time to evaluation" in tables 4a-b-c. This time comprises the wall-clock time to run the experiments, the wall-clock time to run the diagnostics, and the human time needed for producing the assessment. From the table, it is apparent that the time to evaluation is the main bottleneck for swift implementation of cycles, for the seasonal application.

Assessing the skill of the seasonal forecast is troublesome because obtaining significant scores (let alone score differences) at mid-latitudes requires many ensemble members. This makes testing the seasonal range at full resolution difficult. The task is highly parallel, and with enough nodes and disk space in principle any experiment could be run overnight. In practice, any single experiment is usually limited to about 10% of total RD usage. Therefore, a single experiment can take some weeks to complete, and typically only a single experiment can be run at a time.

The Tier (C) set-up can be used for the evaluation of ENSO forecasts, ENSO teleconnections, and some aspects of the model mean state and variability. This evaluation is conducted for each cycle, and it is relatively agile although the diagnostics are not well distributed. Tier (B) is used prior to operational implementation, to obtain guidance on the skill of user-relevant variables such as 2m-temperature and precipitation. Tier (B) takes a considerable amount of both CPU time, wall-clock time and human time, resulting in a poor turnaround of results. More agile and distributed diagnostic software will facilitate

evaluation, and allow model cycles to be evaluated from the seasonal perspective prior to operational implementation. But diagnostic software is not the only limiting factor; there are other scientific challenges, that will be discussed in the next sections.

Tier (A) Operational reforecast set Reforecasts with full configuration and a-posteriori skill assessment		
	ENS monthly forecast range	**Seasonal forecast range**
Reforecast	- Last 20 years - 11 members - Twice weekly - ENS resolution - 46 day forecast length	- 36 years (1981-2016) - 25 members - Monthly - Seasonal resolution - Seasonal forecast length
Cost	~ 33,000 MRE per year (continuous production)	~ 25,000 MRE (spread over a few months, generated every 4-5 years, when the seasonal ensemble is upgraded)
Evaluation criteria	- Key processes (MJO, teleconnections) - Bias and skill scores - Score card	- Key processes (ENSO, teleconnections) - Mean, variability and skill scores - No consolidated scorecard

Table 4a: Summary of the Tier (A) operational reforecasts configuration, their cost and the evaluation criteria applied to assess forecast quality, for the ENS monthly forecast range (column 2) and the seasonal forecast range (column 3). All the integrations are conducted with the coupled model. The unit for the cost is a medium-range-equivalent (MRE): the cost of a single 15-day coupled forecast member. See text for details on the evaluation criteria.

Tier (B) Pre-implementation Reforecasts with reduced temporal sampling		
	ENS monthly forecast range	**Seasonal forecast range**
Reforecast	- 28 years (1989-2016) - 15 members - 4 times per year (Feb, May, Aug, Nov starts) - ENS legB resolution from day 0 - 46 day forecast length	- 36 years (1981-2016) or 24 years (1993-2016) - 25 members - Twice a year (May, Nov starts) - Seasonal resolution - Seasonal forecast length
Cost	~ 1,100 MRE per year	~ 4,800 MRE (or 2,800 MRE for 23y)
Evaluation criteria	- Key processes (MJO, teleconnections) - Bias and skill scores - Score card	- Key processes (ENSO, teleconnections) - Mean, variability and skill scores - No consolidated scorecard
Time to evaluate	~ 2 weeks	~ 2-3 months

Table 4b: As Table 4a but for the Tier (B) pre-implementation reforecasts configuration with reduced sampling.

Tier (C) Specific Developments Reforecasst with reduced temporal sampling, lead-time, number of ensemble members and resolution		
	ENS monthly forecast range	**Seasonal forecast range**
Reforecast	- 28 years (1989-2016) - 15 members - 4 times per year (Feb, May, Aug, Nov starts) - ENS legB resolution from day 0 - 32 day forecast length	- 36 years (1981-2016) or 23 years (1993-2016) - 5 members - Twice a year (May, Nov starts) - Reduced resolution - Seasonal forecast length
Cost	~ 730 MRE per year	~ 130 MRE
Evaluation criteria	- Key processes (MJO, teleconnections) - Bias and skill scores - Score card	- Key processes (ENSO, teleconnections) - Mean, variability and skill scores - No consolidated scorecard
Time to evaluate	~ 2 weeks	~ 2 weeks

Table 4c: As Table 4a but for the Tier (C) specific development reforecasts configuration with reduced sampling, lead-time, number of ensemble members and resolution.

Appendix B – Tables of tests: present and proposed changes

Alpha-0: Early cheap technical tests

No change is proposed.

Alpha-1a: Individual component tests

The second layer of tests are individual component tests, where for the first time we try to gain insight into the meteorological impact of changes. These tests are more expensive than the technical ones, but still need to be affordable as many need to be run. It is particularly at this stage that several difficult compromises are needed, as its unaffordable to closely replicate the operational configuration, yet it is important to have as accurate as possible an evaluation of the merit of the change.

Alpha-1a - Individual component tests	
Component being evaluated	**Test**
HRES & 4D-Var	- At present ocean is run at ¼ degree resolution and atmosphere TCo399. *- Proposed change - 1-degree ocean for a 6-month summer and winter (subject to further work to confirm validity of this approach).*
EDA	- At present no early assessment of EDA is made. *- Proposed change - For DA, EDA and some observation changes only: reduced member and resolution slim EDA to assess spread for changes where an impact could be reasonably expected. One month of summer and winter tests should be enough to highlight any sensitivities. In addition to monitoring the EDA variance and B estimate, the performance of the EDA control member can be assessed.*
ENS medium range	- At present, in general, early assessment of ENS is now often made, although this is a recent innovation (and configurations/use not yet optimized). *- Proposed change - For changes originating from the Ensemble Perturbation Methodology area or where an interaction has been identified or could be reasonably expected (e.g. model physics): Reduced resolution and ensemble size ENS started from TCo399 deterministic analysis experiment to assess changes in spread and probabilistic skill.*
ENS monthly extension	- At present, in general, early assessment of ENS is made for many model changes *- Proposed change - For specific changes in model, ensemble: 5-member ensemble of re-forecasts at a reduced ocean (1 degree, 75 vertical levels) and atmosphere resolution (Tco319) initialized from ERA5, starting the 1st of each month over the period 1989-present.*
Seasonal	- At present, there is no test of scores. For specific model changes, seasonal suite is used to assess model climate. *- Proposed change - Enhanced efforts to assess impact of model changes on climate.*
CAMS	- At present no test, except tests of CAMS changes done by CAMS. *- No proposed change*.
Re-analysis	- At present no test. *- No proposed change.*

Table 5: Individual Component stage testing requirements.

Alpha-1b: Thematic merge tests

Once the individual changes have been tested our process is to test merged sets of changes, to create thematic merges that can be merged in sequence in the following phase.

Alpha-1b - Thematic merge tests	
Component	**Test**
HRES & 4D-Var	- At present for all thematic merges: Coupled TCo399 atmosphere and ¼ degree ocean for a 6-month summer and winter. *- No proposed change.*
EDA	- At present no early assessment of EDA is made. *- Proposed change - Reduced member and resolution slim EDA to assess spread for changes where an interaction could be reasonably expected. These tests will ensure that the EDA statistics change consistently (e.g. EM errors vs ES spread) together with an assessment of the performance of the EDA control member. One month of summer and winter tests should be enough to highlight any sensitivities. An open question is whether this can be extended in future to run 6 months and use the B and variance estimates in the TCo399 HRES-like testing. However, we need to assess if this is feasible or not. We could test this after the 46R1 testing is complete (run experiments in parallel to the TCo399 testing).*
ENS medium range	- At present ENS evaluation is not consistently undertaken at this stage, although some early evaluation is sometimes done. *- Proposed change - Model and ENS thematic merges: Reduced resolution and ensemble size ENS started from 6-month summer/winter TCo399 HRES analysis experiment to assess changes in spread and probabilistic skill*
ENS monthly extension	- At present no test. *- No proposed change.*
Seasonal	- At present no test *- No proposed change*.
CAMS	- At present no test other than tests done by CAMS contributors. *- No proposed change.*
Re-analysis	- At present no test *- No proposed change*.

Table 6: Merged component stage testing requirements.

Alpha-2: Research Department merge tests

Once the individual changes have been tested our process is to test merged sets of changes, building an RD e-suite sequentially.

Alpha-2 - Research Department merge tests	
Component	**Test**
HRES & 4D-Var	- At present for all RD merges: Coupled TCo399 atmosphere and ¼ degree ocean for a 6-month summer and winter. *- No proposed change.*
EDA	- At present EDA testing is only done at the very final stage of RD merge tests. *- Proposed change - All RD merges: Reduced member and resolution EDA to assess spread. These tests will ensure that the EDA statistics change consistently (e.g. EM errors vs ES spread) together with an assessment of the performance of the EDA control member. One month of summer and winter tests should be enough to highlight any sensitivities.*
ENS medium range	- At present, ENS testing is now routinely done at this stage, although this is a recent innovation (and configurations/use not yet optimized). *- Proposed change - All RD merges: Reduced resolution and ensemble size ENS started from 6-month summer/winter TCo399 HRES analysis experiment to assess changes in spread and probabilistic skill*
ENS monthly extension	- At present a test is run using a 5-member ensemble of re-forecasts at a reduced atmosphere resolution (Tco319) initialized from ERA5, starting the 1st of each month over the period 1989-present. *- Proposed change - For specific changes in the ocean/land/sea-ice data assimilation we need to evaluate consistency with reforecast: the impact can be measured using a reduced set of cases, with 25-51 ensemble members.*
Seasonal	- At present, limited testing. *- Proposed change - Seasonal runs routinely used to evaluate model climate and ENSO scores: 5-10 ensemble members, reduced resolution (Tco199, ORCA1-75 levels); use of stochastic physics; Nov and May starts from 1981-2016. Additionally, for specific changes in the ocean/land/sea-ice data assimilation we need to evaluate consistency with reforecast: the impact can be measured using a reduced set of cases at full resolution, with 25-51 ensemble members.*
CAMS	- At present no test, other than tests done by CAMS contributors. *- No proposed change.*
Re-analysis	- At present no testing is done specific to re-analysis. *- Proposed change - All RD merges: Ensure that the infra-structure for running IFS in the past is not broken (C3S). Assess changes in the free model climate. Consider testing with reduce observation set*.

Table 7: Merged RD stage testing requirements.

Alpha-3: RD e-suite

Once the full merged cycle is agreed we run a full RD e-suites before handover to FD

Alpha-3 - Research Department e-suite tests	
Component	**Test**
HRES & 4D-Var	- At present a coupled TCo1279 atmosphere and ¼ degree ocean for a 6-month summer and winter, including early delivery is run. *- No proposed change.*
EDA	- At present a full EDA testing for a 6-month summer and winter, including early delivery is run. *- No proposed change.*
ENS medium range	- At present a full ENS testing for a 6-month summer and winter, including early delivery is run *- No proposed change.*
ENS monthly extension	- At present the RD e-suite is tested in the extended range with a 5-member ensemble of re-forecasts at a reduced atmosphere resolution (Tco319) initialized from ERA5, starting the 1st of each month over the period 1989-present. *- No proposed change*.
Seasonal	- At present a full evaluation is undertaken only for cycles to be used for new SEAS version. *- Proposed change - Even for cycles not to be used for new SEAS version, once the e-suite is stable, a full evaluation with 25 ensemble members at operational resolution, 1981-2016, May and November starts, with stochastic physics. If necessary, assessment can be completed after operational implementation of cycle.*
CAMS	- At present no test done. *- No proposed change, but this remains an open issue and should be revisited.*
Re-analysis	- At present no test is run. *- Proposed change - Ensure that the infra-structure for running IFS in the past is not broken (C3S). Assess changes in the free model climate. Consider testing with reduce observation set.*

Table 8: RD e-suite test requirements.

Appendix C – List of acronyms

- CAMS: the Copernicus Atmospheric Monitoring Composition Service;
- CRPS: the Continuous Rank Probability Score;
- C3S: the Copernicus Climate Change Service;
- EDA: the ECMWF ensemble of data assimilations (it includes 25 members run at Tco639L137 resolution);
- ENS: the ECMWF medium-range/monthly ensemble (it includes 51 members, run with a Tco639L91 resolution up to forecast day 15, and Tco319L91 from day 15 to 46);
- ENSO: El Nino Southern Oscillation;
- FD: Forecast Department;
- FGAT: first-guess at the right time, a characteristic of the NEMOVAR 3d-Var data assimilation system;
- HRES: the ECMWF (single) high-resolution (Tco1279L137) forecast;
- HRES-4DV: the ECMWF high-resolution (now Tco1279L137) assimilation;
- HPC: High Performance Computer;
- IFS: the ECMWF Integrated Forecast System;
- IWG: the ECMWF Implementation Working Group, established to develop, test, and implement each model cycle;
- LIM: the Louvain-la-Neuve sea-ice model;
- LWDA: the ECMWF Long Window (12 hour) 4-dimensional variational Data Assimilation run twice a day, at 00 and 12 UTC
- LNNN: vertical resolution, expressed in terms of the total number of vertical levels;
- MJO: the Madden-Julian Oscillation;
- NAO: the North Atlantic Oscillation;
- NEMO: the Nucleus for European Modelling of the Ocean;
- OCEAN4: the ECMWF ocean analysis version 4, with a 1°-degree horizontal resolution and 42 vertical levels;
- OCEAN5: the ECMWF ocean analysis version 5;
- ORAS4: the ECMWF Ocean reanalysis version 4 (it included 5 members, run with a 1°-degree horizontal resolution and 42 levels NEMO ocean model);
- ORAS5: the ECMWF Ocean reanalysis version 5 (it includes 5 members, run with a 0.25°-degree horizontal resolution and 75 levels NEMO ocean model, and LIM sea-ice model);

- ORCA: the NEMO tri-polar grid;
- PNA: the Pacific North-Atlantic pattern;
- QED: the ECMWF Quarterly Evaluation and Development meeting;
- RD: Research Department;
- RMSE: Root-Mean-Square Error;
- ROCA: the area under the Relative Operating Characteristic curve, a measure of a probabilistic system capability to discriminate between occurrence and non-occurrence of dichotomous events;
- R2O: the Research-to-Operations process;
- SEAS: the ECMWF seasonal ensemble;
- SEAS4: the ECMWF seasonal ensemble, version-4 (it includes 51 members run with a TL255L91 resolution, up to 7 months; extended to 13 months every quarter);
- SEAS5: the ECMWF seasonal ensemble, version-5 (it includes 51 members run with a Tco319L91 resolution, up to 7 months; extended to 13 months every quarter);
- SEEPS: the Stable Equitable Error in Probability Space, used to assess the quality of single, precipitation forecasts;
- SST: Sea-Surface Temperature;
- S2S: the WMO WWRP/WCRP Subseasonal-to-seasonal project;
- TCoNNN: spectra triangular truncation with NNN total wave numbers, and cubic-octahedral grid in physical space;
- TLNNN: spectral triangular truncation with NNN total wave number, and linear grid in physical space;
- VarBC: Variational Bias Correction;
- 3d-Var: 3-dimensional variational data assimilation system;
- 4d-Var: 4-dimensional variational data assimilation system;
- 4DVAR: acronym used to denote the ECMWF high-resolution (now Tco1279L137), 4d-Var data assimilation system.