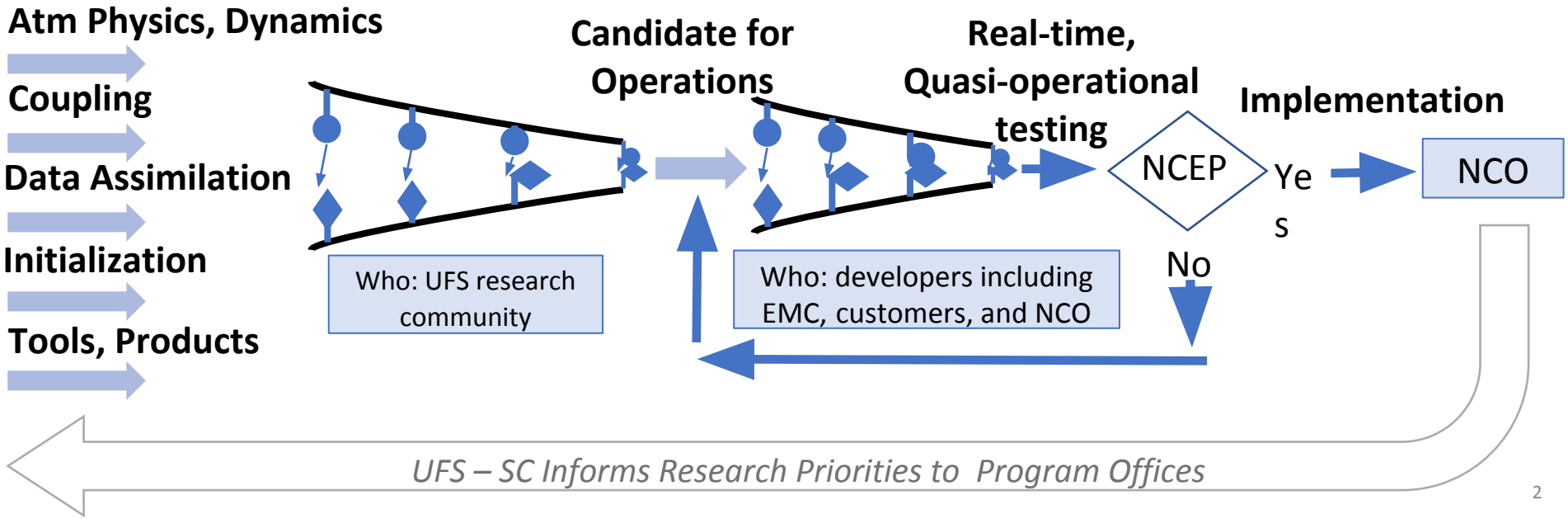
Test plan for GFS

R202R: Improving by Doing

- **Testing and evaluation are critical activities for R2O process. They provide the evidence for decisions to pass through the gates. Formalize them in a test plan.**
- **Prioritize testing goals, combine community and operational knowledge and tools, and adopt best practices to develop and select best possible prediction system.**



Framing questions and concepts for the test plan

- How do we establish what matters?
  - Evaluation priorities
  - Targets (e.g. absolute error limit, relative improvement)
  - Decision under mixed results (e.g. score cards, indices)
- What are community best practices?
- Should we adopt the concept of spiral development: evaluate, improve, evaluate again?
  - Scope of evaluation in different stages, process-based diagnostics
- Hierarchy of end-to-end system evaluation
- How to establish minimal and optimal testing needs?
- Automation to support evaluation
  - Selection of standard tests and metrics
  - Inclusion in workflow

Goals of operational system development cycle

1. **Meet requirements and be responsive to stakeholder input** (e.g. large scale flow, high impact events)
2. **Address identified shortcomings of the current system** (e.g. from MEG presentation: cold bias, low level inversions)
3. **Incorporate relevant science improvements** (e.g. PBL physics parameterization, component coupling)
4. **Advance strategic goals** (e.g. unified forecast system development)

Considerations

- **Scope:**
  - Determined by improvement goals or by schedule
  - Maturity/refinement of scientific algorithms

- **Implementation schedule**:
  - Computing necessary for testing and retrospectives
  - Moratorium on implementations due to computer system change (expected for Q3 FY21- Q2 FY 22)
  - Scope of evaluation - length of retrospectives and real-time parallels

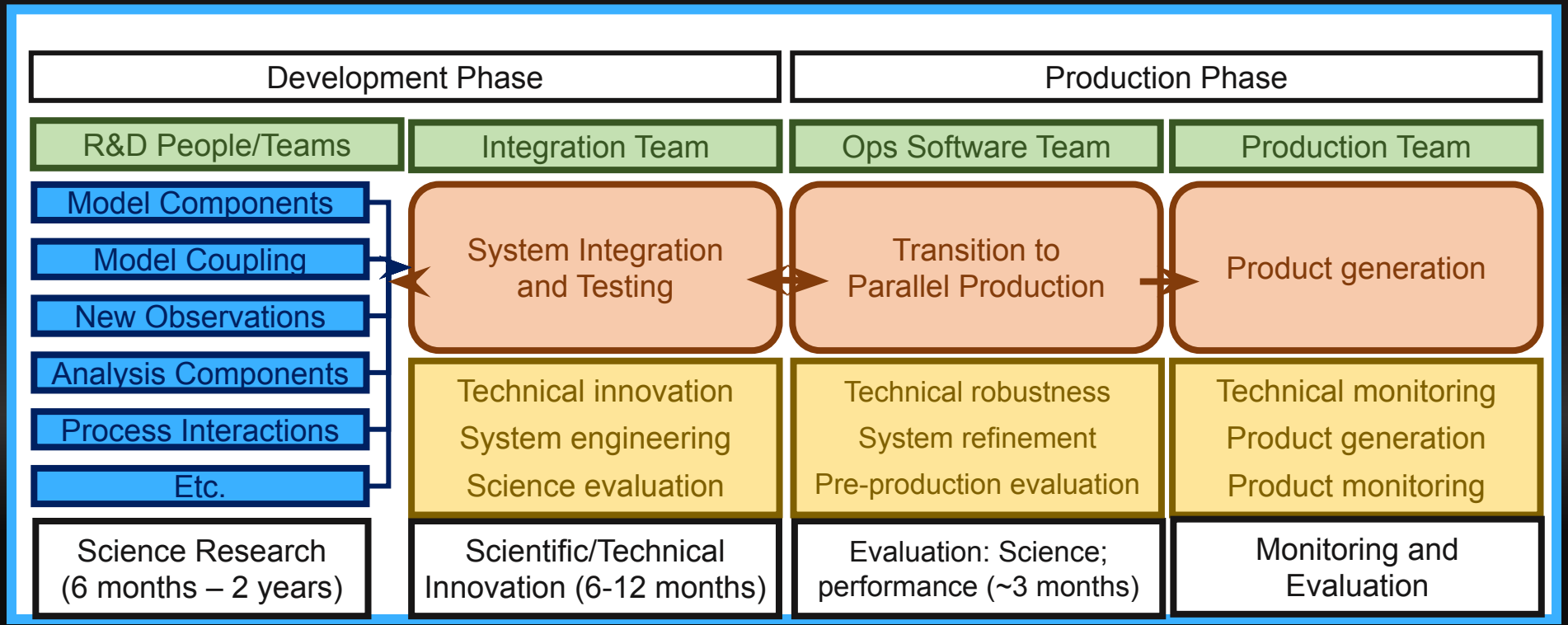
Examples of community practices for testing and evaluation

Hierarchical System Development under consideration can inform testing design

- Ability to test atmospheric physics using a single column model (SCM).
- Ability to turn model component feedbacks off using data components.
- Ability to use model component configurations that have simplified scientific algorithms.
- Ability to run a model with pre-assimilated initial conditions and cycling, but no data assimilation.
- Ability to run a quasi- or full operational workflow, including data assimilation.

https://docs.google.com/document/d/1A9upcMYvYjz_gj2TPuZhoNwvTT8fCTP8qi02djWmlF8/edit

Development Pathway for GEOS Systems From Research to Production



Courtesy: Steven Pawson, GMAO

ECMWF testing hierarchy

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:
  - o 0 - Individual ad-hoc testing — Developer's branch on HPC with PrepIFS; build IFS locally and run a small suite of model test
  - o 1a - Individual testing against controls — Control = current ops at reduced res., few ens. members
  - o 1b – Thematic pre-merging and addressing interactions — Combine changes that interact into suites. Test vs. reduced res. control
  - o 2 - Incremental build and testing — Versions of increasing completeness. Summer & winter tests against reduced res. control. ENS from ops analysis every 8 days
  - o 3 - RD e-suite: higher resolution HRES/ENS and EDA testing — Full resolution, summer & winter
- The Beta-phase testing; — Experimental suite in parallel with current operational suite. Few model changes and they are well documented.
- The Release-Candidate-phase testing. — Frozen code, all model products provided

IFS at reduced resolution

Buizza et al., ECMWF Tech Memorandum [829]

ECMWF holistic evaluation

- Headline scores for the operational forecasts;

500 hPa ACC, 850 hPa T, precip, extreme forecast index, hurricane track and intensity, 2m T

What

- 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;

How

- 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);

Metrics

- 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

- Technical evaluation (computational cost, memory usage, code refactoring).

Buizza et al., ECMWF Tech Memorandum 829

V&V WG T&E Recommendations

General

- Consider ECMWF-like testing cycle: Alpha-phase, Beta-phase, and Release-candidate-phase testing
- Metrics and diagnostics need to identify strengths and weaknesses and allow the developers to determine where to look for improvements
- Once weaknesses are identified, select additional metrics to measure what we are trying to “fix” and “maintain”
- Suite of metrics should be complimented by subjective evaluations
- Possible way to define metrics: holistic categories such as large-scale flow, high-impact weather, tropical cyclones, etc ....
- Coupling evaluation needs knowledge of climatologies through reforecasts

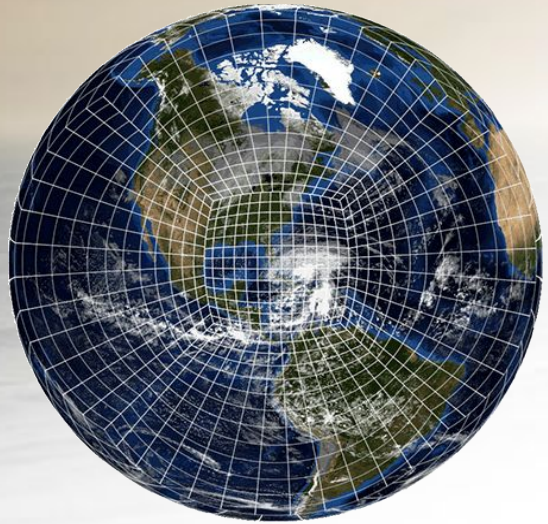
Community Involvement

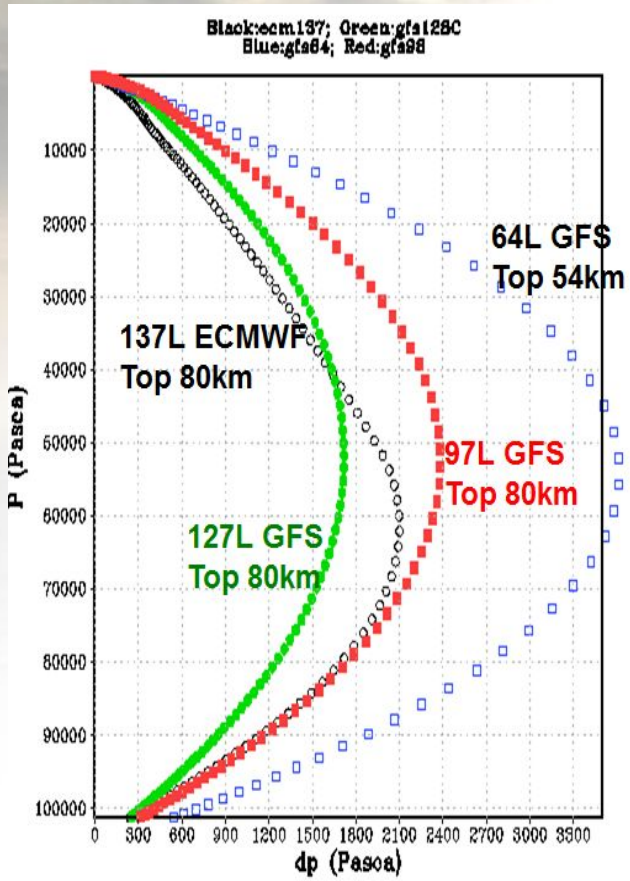
- Work with universities, private sector, other NOAA entities to get more “eyes on” operational vs parallels runs for evaluation
- Publish test plans well in advance so community can identify areas of evaluation not covered by EMC where they can contribute
- Need methods of synthesizing metrics and scorecards
- Need more focus on observation data sources for independent evaluation

GFS/GDAS upgrades

GFS/GDAS v15.1

- FV3 dynamical core and other upgrades





GFS/GDAS v16

- Configuration
  - Increased resolution
  - Advanced Physics
  - Coupled to Wave Model
  - Improved Data Assimilation
- Planned for implementation in Q2FY21

Potential Major Scientific Upgrades for GFS V16

Model

- **Domain and resolution:**
  - Increased vertical resolution from 64 to 127 vertical levels and raise model top from 54 km to 80 km; Increased horizontal resolution from 13 km to 10 km (depending on operational resources)
- **Dynamics:** New advection algorithms from GFDL
- **Advanced physics chosen from Physics Test Plan:**
  - PBL/turbulence: K-EDMF => sa-TKE-EDMF
  - Land surface: Noah => Noah-MP
  - Gravity Wave Drag: => unified gravity-wave-drag
  - Radiation: updates to cloud-overlap assumptions,
  - Microphysics: Improvements to GFDL MP
- **Coupling to WaveWatchIII**
  - Two-way interactive coupling of atmospheric model with Global Wave Model (GWM)

Data Assimilation:

- Local Ensemble Kalman Filter (LETKF), including early cycle updates in support of GEFS
- 4-Dimensional Incremental Analysis Update (4DIAU)
- Stochastic Kinetic Energy Backscatter (SEKB) based land surface perturbations
- Stratospheric humidity increments
- Improved Near Surface Sea Temperature (NSST) analysis
- Land Data Assimilation
- Shifting and Lagging Ensemble Members to expand ensemble size
- Improved cloud analysis
- Delz increments

GFS - physics testing

J. Kain https://ufscommunity.org/docs/Repository/20190501_GFS_Physics_Suite_Testing_Report.pdf and references therein

	Suite 1 (GFS v15)	Suite 2	Suite 3	Suite 4
Deep convection	sa-SAS	sa-SAS	sa-CSAW	sa/aa-GF
Shallow convection	sa-MF	sa-MF	sa-MF	MYNN-EDMF and sa GF
Microphysics	GFDL	GFDL	aa-MG3	aa-Thompson
PBL/Turbulence	K-EDMF	sa-TKE-EDMF	K-EDMF	MYNN-EDMF
Land Surface Model	Noah	Noah	Noah	RUC

Table 1. Physics suites evaluated for possible implementation in GFSv16.
sa: Scale-aware; aa: aerosol aware; SAS: Simplified Arakawa Schubert; MF: Mass flux; MYNN: Mellor–Yamada–Nakanishi–Niino; EDMF: Eddy-diffusivity/Mass-flux; TKE: turbulent kinetic energy; CSAW: Chikira-Sugiyama-Arakawa-Wu; GFDL: Geophysical Fluid Dynamics Laboratory; MG3: Morrison-Gettelman; RUC: Rapid Update Cycle.

sts at C768L64 initialized from
very five days between
12/31/2017, alternating between
C, plus 16 MEG-identified cases
clone (TC) cases plus 8 other:

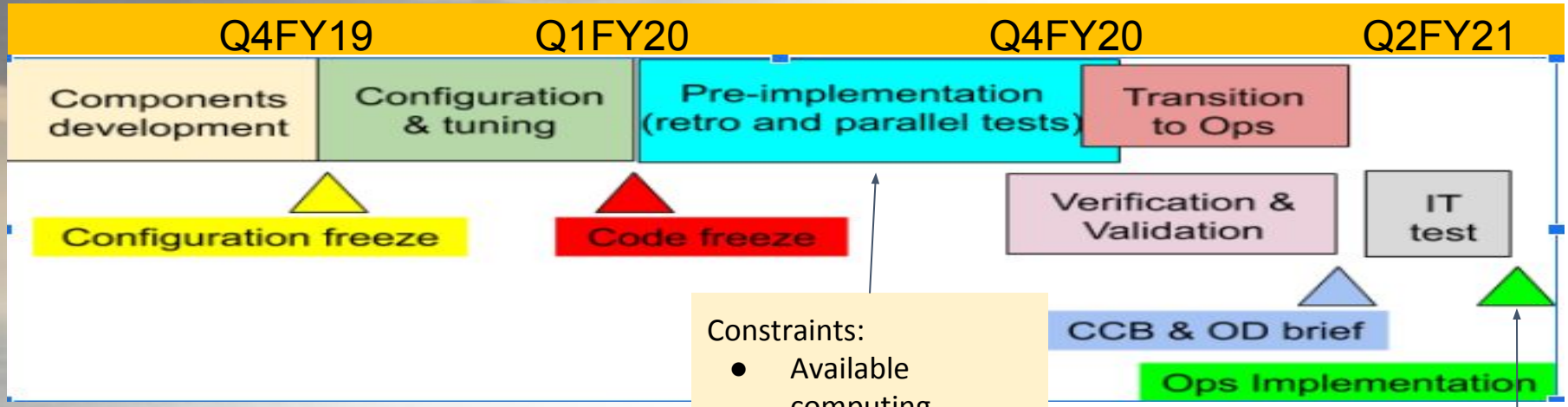
- Blizzard of 2016 - progressive
- z Plains severe weather - progressive,
- ce to examine drylines
- z "Pi Day" Blizzard - Precipitation type
- 4/20/17 00Z Flooding in the Mississippi Valley
- 7/29/17 00z Too hot in FV3GFS in CA
- 10/16/17 12z Inversions and 2-m temperature
- 1/1/18 00z "Bomb" cyclone
- 3/15/17 00z Atmosphere river - progressive

- **PBL/turbulence:** K-EDMF => sa-TKE-EDMF
- **Land surface:** Noah => Noah-MP
- **GWD:** separate orographic/non-orographic => unified gravity-wave-drag
- **Radiation:** updates to cloud-overlap assumptions, empirical coefficients, etc. in RRTMG

These upgrades will be introduced along with a much higher model top and up to twice as many vertical levels as GFSv15, beginning immediately. Selection of a tuned prototype configuration for GFSv16, including these upgrades and an updated data assimilation package, is expected to be completed by the end of FY19.

Evaluated ACC, RMSE, bias, upper air T, 2m T, CONUS precipitation, TC track and intensity, PBL inversions, cases

GFS V16 Implementation Schedule



Constraints:
- Available computing,
- Length of retrospective testing

Constraint:
- Start of moratorium

Detailed Project Plan and Charter being developed: Draft version

Hierarchy of end-to-end system evaluation

Using full system targeted for operations, perform hierarchical evaluation of proposed changes.
How to determine minimum required testing and optimal testing needs given computing and human resource considerations e.g. :

- reduced resolution/reduced domain
- target operational horizontal and vertical resolution,

- short forecast-only (up to 1 day),
- forecast-only 10 days and longer (out to seasonal?)
- DA only if DA change
- fully-cycled,

- real-time,
- several canned test-cases for different seasons/phenomena,
- a couple winter and summer months,
- full retrospectives - length?

- Individual changes followed by system integration

- Standardization of testing configurations, metrics, display and synthesis of results

- Automation of T&E

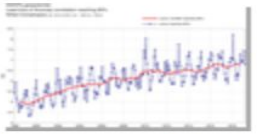
- Spiral development and evaluation

- Timing of subjective evaluation

ECMWF

- Headline scores
- Verification of high-resolution forecasts
- Tracked over long time

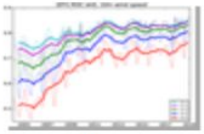
Headline scores

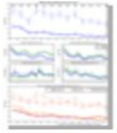

Lead time of ACC reaching a threshold


Lead time of CRPSS of T850


Lead time of CRPSS of 24h


Lead time of 1-SEEPS of 24-h

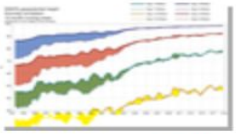

ROC skill score of Extreme Forecast


Errors of tropical cyclone forecasts


Fraction of large 2m temperature errors


RPS of 2m temperature of

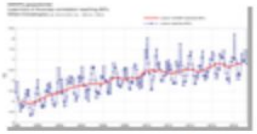
Verification of high-resolution forecasts


Anomaly correlation of ECMWF 500hPa

500 hPa ACC


Verification of the high-resolution

Sensible Wx vs observations


Lead time of ACC reaching a threshold

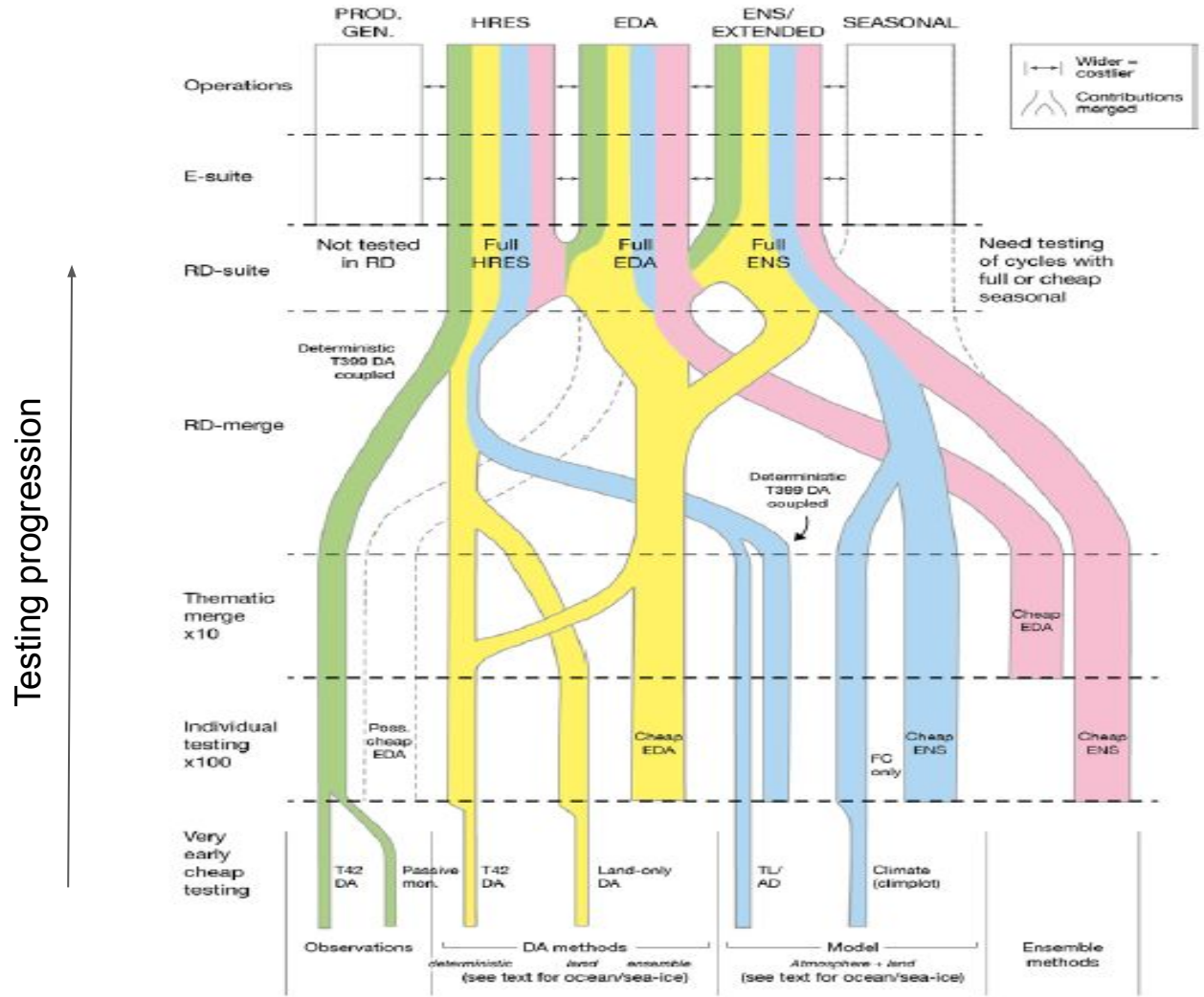
500 hPa ACC


Lead time of 1-SEEPS of 24-h

Precipitation

Schematic of ECMWF testing hierarchical strategy

Width indicates cost



Buizza et al., ECMWF Tech Memorandum 829