1 year of optimizations on the first scalar supercomputer of Meteo-France

P. Marguinaud & R. El Khatib (Météo-France)

24th Aladin Workshop – Hirlam All Staff Meeting 2014

Bucharest 07-10/04/2014

METEO FRANCE
Toujours un temps d'avance

Plan

- Main changes in the software configurations
- Specific optimizations for ARPEGE
- Parallelization of FESTAT
- FULLPOS-2
  - Optimizations
  - The Boyd biperiodicization
- Other enhancements
  - post-processing server
  - FA-LFI
  - IO server

Main changes in the software configurations

Feature	On NEC at M-F	On Bull at M-F
NPROMA	Large value (~3582)	Small value (~50)
MPI gridpoint distribution	North-South only	North-South and East-West
MPI spectral distribution	On waves only	On waves and on vertical levels
LEQ_REGIONS (ARPEGE)	.FALSE.	.TRUE. (except for AROME couplers)
MPI « I/O » distribution (NSTRIN,NSTROUT)	Small values	Maximum values unless I/O server
LOPT_SCALAR	.FALSE.	.TRUE. (to be revisited)
Open-MP	Not used	Used (but not yet in 3DVar)

Specific optimizations for ARPEGE

- **Computation of Gaussian reduced grids :**
  - Merged with ECMWF program
  - Optimized even more for Open-MP
- **Computation of filtering matrixes for the post-processing :**
  - Calculation of dilatation/contraction matrixes :
    - fully recoded to use the algorithm of Strassauber for Legendre polynomial computation
    - Distributed/parallelized
    - Modular
  - Calculation of filtering matrixes themselves
    - Distributed/parallelized
    - Interfaced with the modular calculation of dilatation/contraction matrixes

=> *On-line recomputation of filtering matrixes is fast.*

No need anymore for huge matrixes on files.

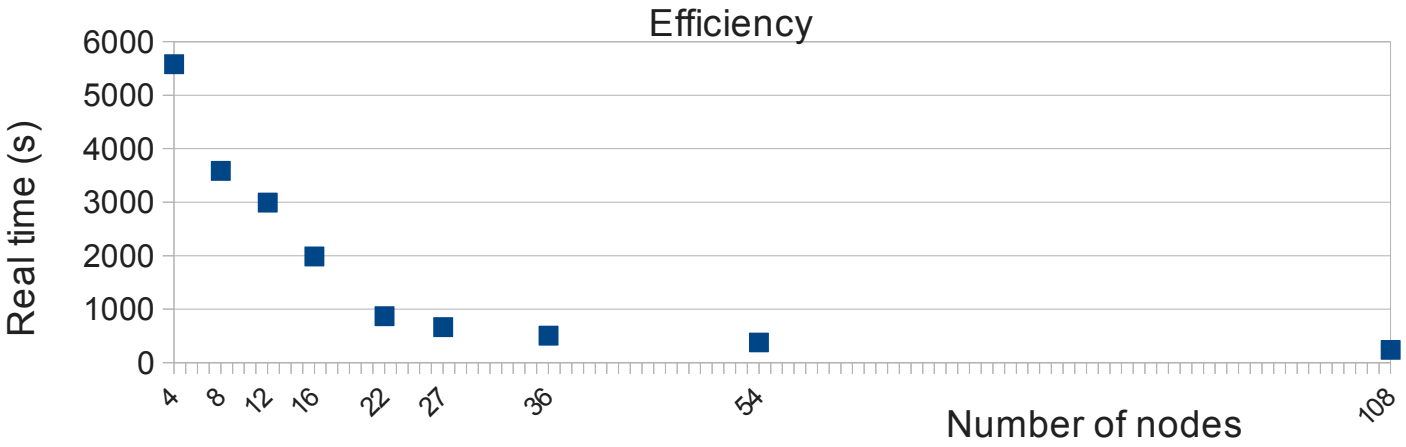
Parallelization of FESTAT (1)

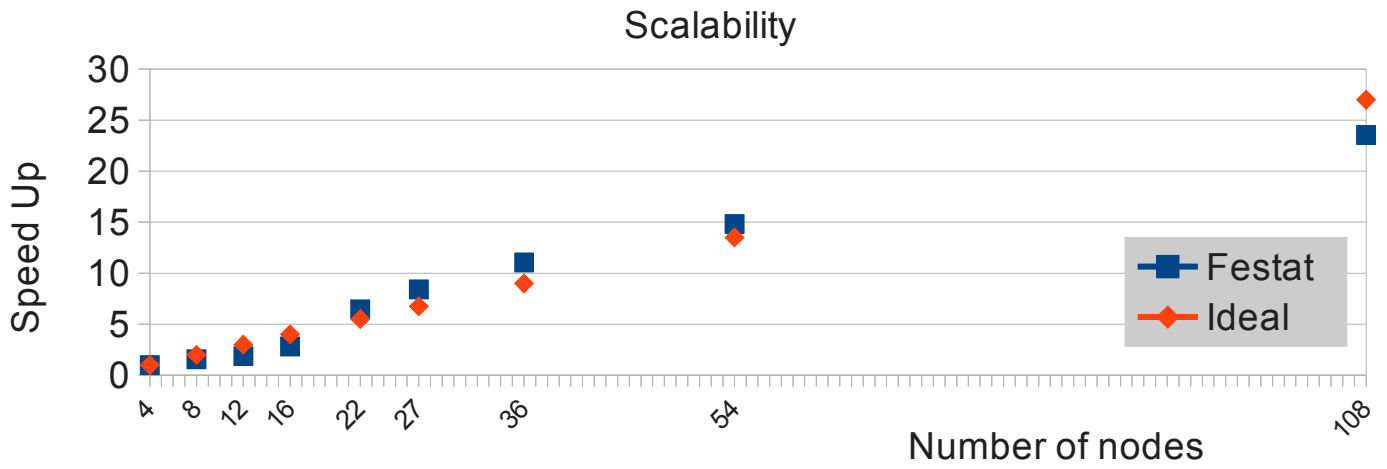
FESTAT : Forecast Error STATistics :

a mono-task program developped during the years 1993-1997 which reads a lot of files and makes a lot of summations

- Substantial re-write and cleaning, DrHook instrumentation
- Optimizations by inverting loops and re-organization of calculations in order to limit repeated computations
- Memory savings
- Development of an « in-core » option to read the files only once
- MPI distribution on the files => needs to re-order the summation independently of the members for bitwise reproducibility
- MPI distribution + Open-MP parallelization on the k^* wave numbers
- MPI distribution on the m waves is possible but not coded yet

Parallelization of FESTAT (2)





108 members, 90 levels, truncation 719 x 767

72 nodes (54 tasks x 12 threads) => 11 min.

METEO FRANCE
Toujours un temps d'avance

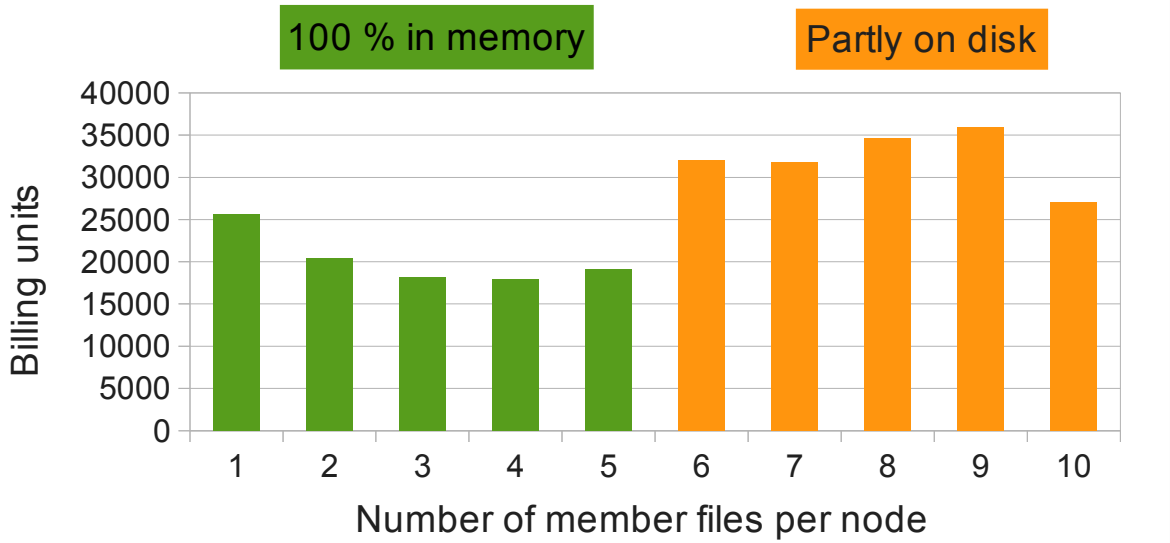
Parallelization of FESTAT (3)

Getting started with parallel FESTAT :

OMP_NUM_THREADS=$(number of threads)

Namelist file NAMFESTAT in fort.4

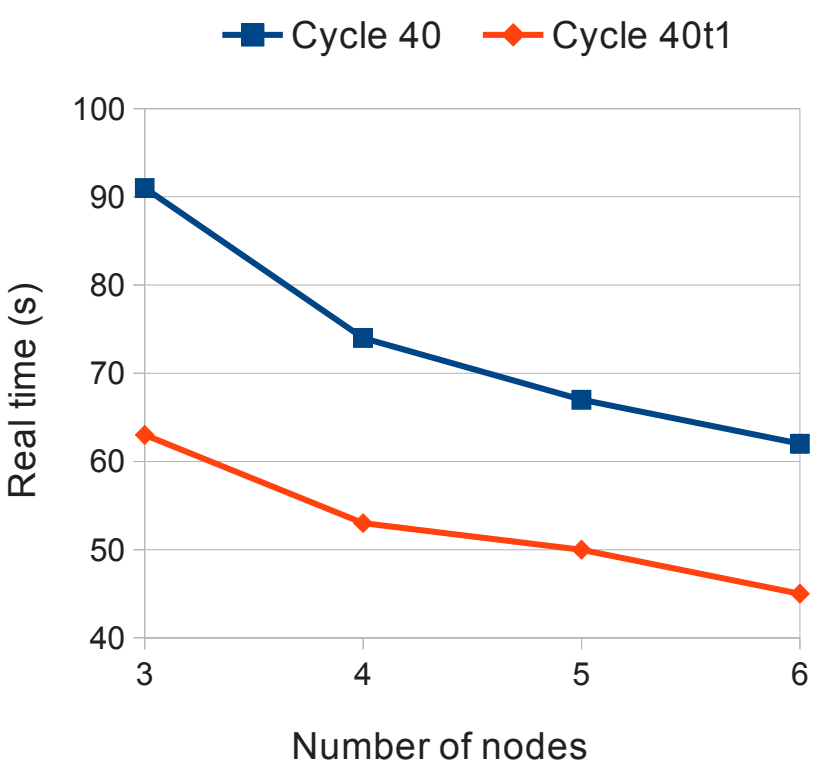
mpirun -np $(number of mpi tasks) FESTAT

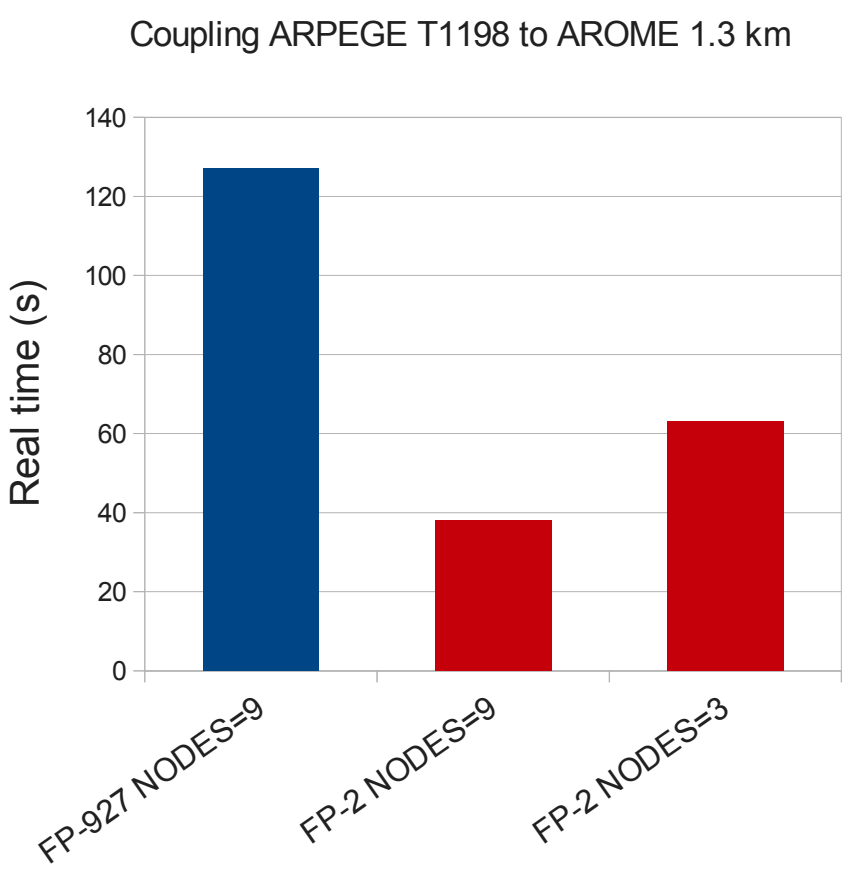


=> read festat_guidelines.pdf for more

&NAMFESTAT
NFLEV=60,
NCASES=72,
LSTABAL=.true.,
LANAFBAL=.true.,
OUTBAL='stabbal',
OUTCVT='stabcvt',
OUTCVU='stabcv',
ELON1=351.613266296832,
ELAT1=37.33305050,
ELON2=15.74536410,
ELAT2=53.02364987,
ELON0=2.00000000,
ELAT0=45.80000000,
NMSMAX=374,
NSMAX=359,
NDGL=720,
NDLON=750,
NDGUX=709,
NDLUX=739,
EDELY=2500.00000000,
EDELX=2500.00000000,
LELAM=.TRUE.,
/

Optimization of Fullpos-2





The Boyd biperiodicization in Fullpos-2 (1)

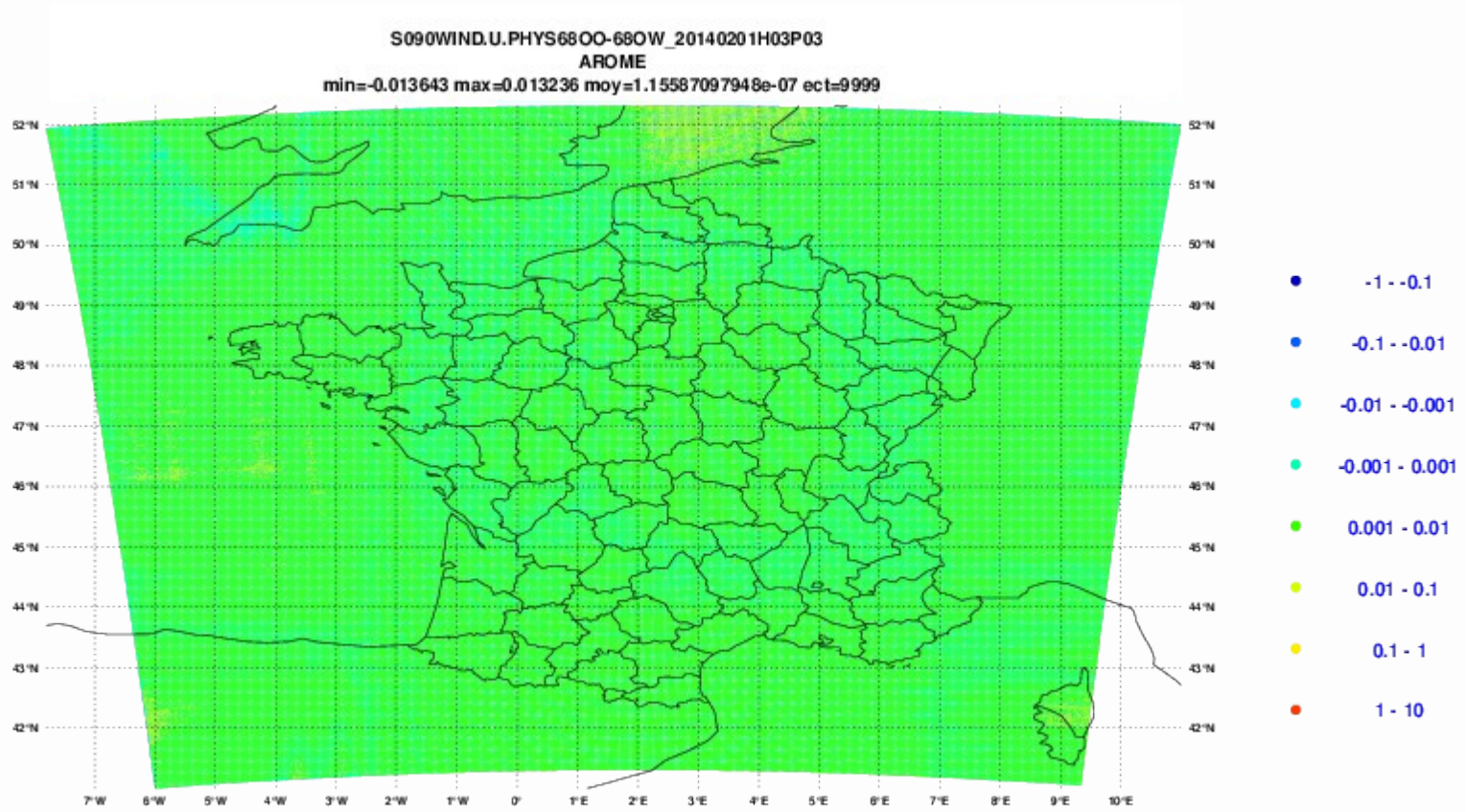
- In New developments have been added :
  - The interpolation grid and the target grid can be different
  - **=> Needs to move the biperiodicization immediately after the horizontal interpolation**
  - **=> Makes less MPI communications**
  - **=> Biperiodicization code is easier to maintain**
  - **=> Switching from splines to Boyd biperiodicization becomes easy :**

    just set in namelist NAMFPC:

    NFPBOYD=1

The Boyd biperiodicization in Fullpos-2 (2)

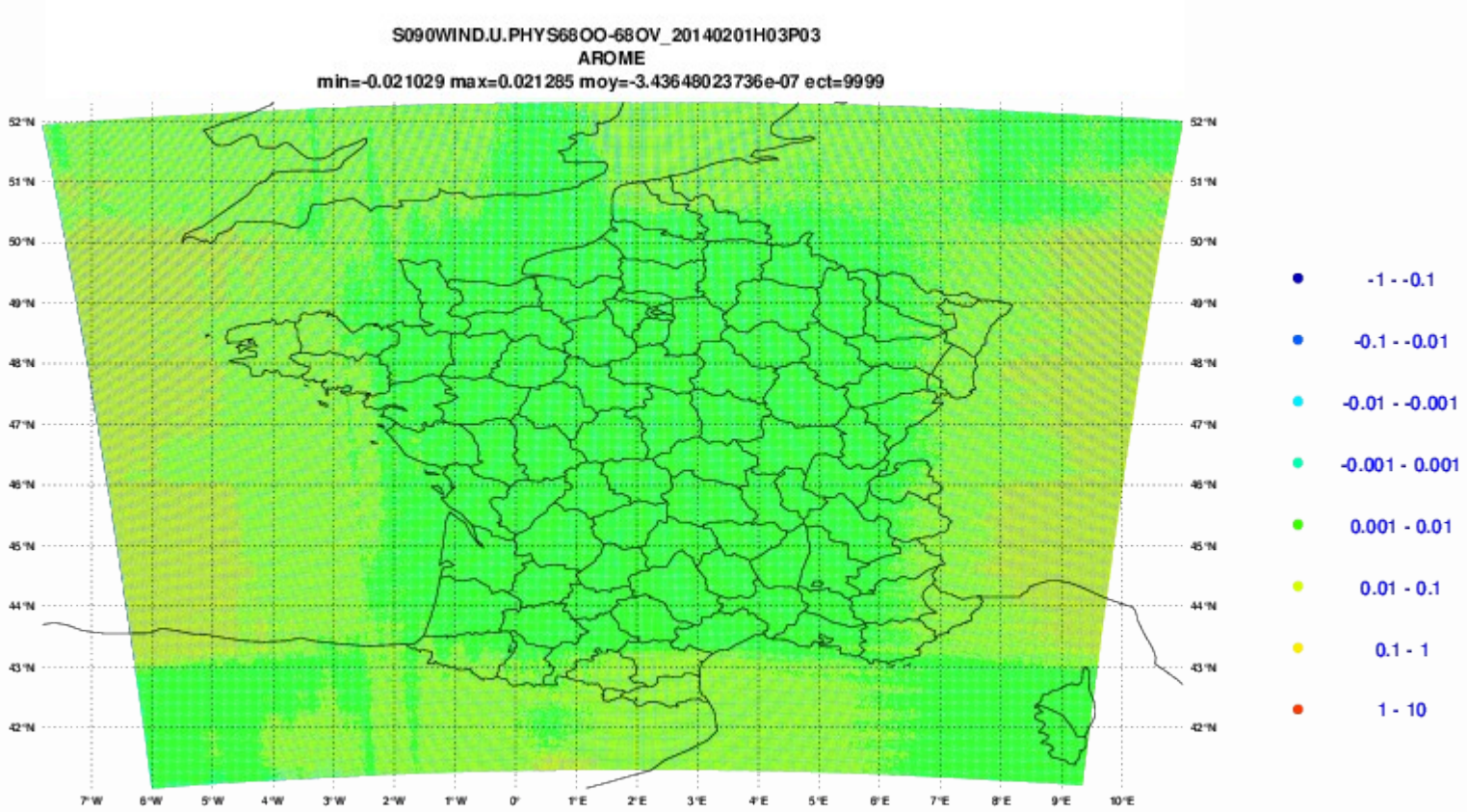
Difference Splines/Boyd on the U-wind at the lowest level



METEO FRANCE
Toujours un temps d'avance

The Boyd biperiodicization in Fullpos-2 (3)

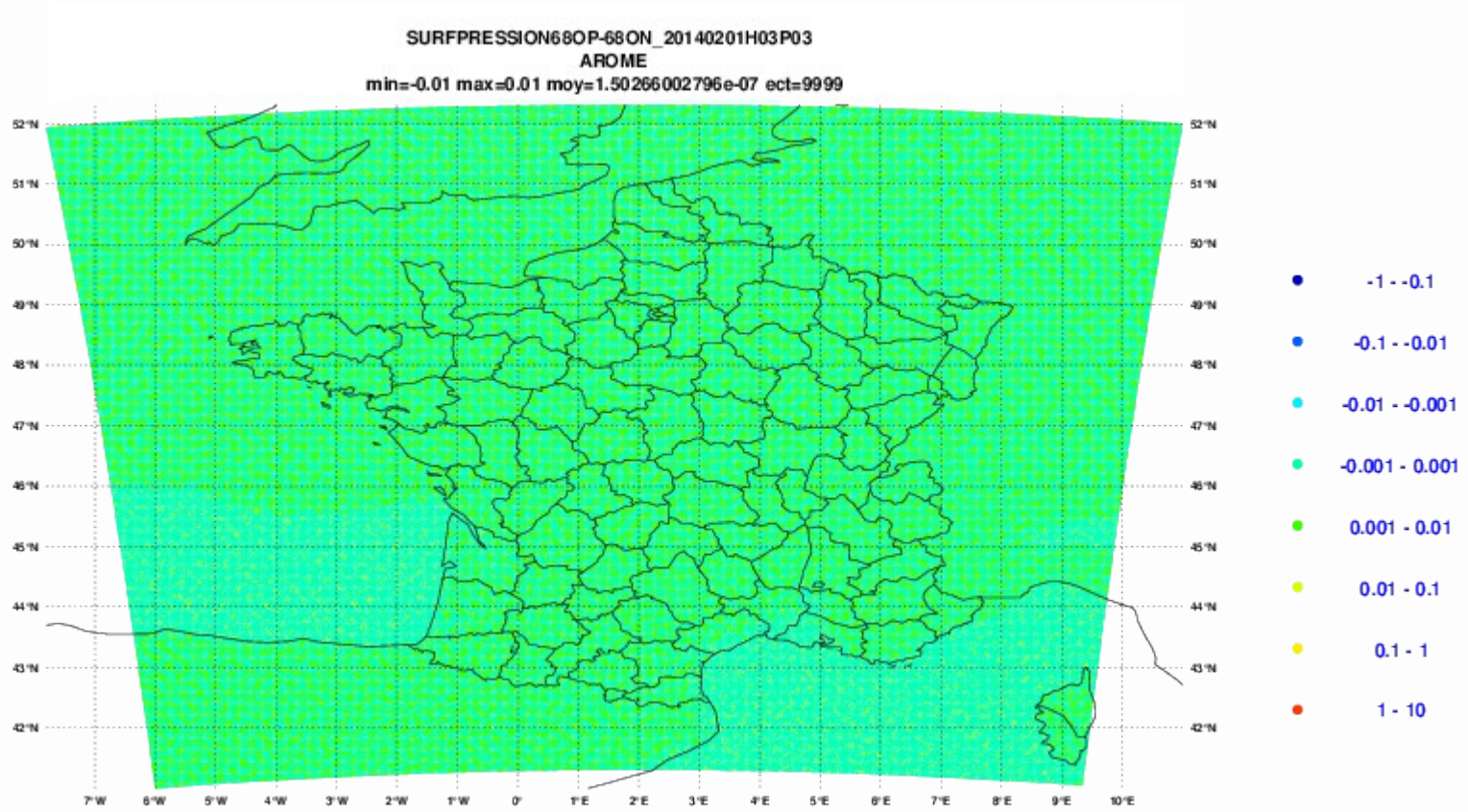
Difference of the biperiodicization Before/After the vertical interpolations on the U-wind at the lowest level



METEO FRANCE
Toujours un temps d'avance

Fullpos-927 will no longer be supported in cycle 41

Difference between NFPOS=927 and NFPOS=2
on surface pressure



METEO FRANCE
Toujours un temps d'avance

Post-processing server

- Start the model once, post-process multiple files (same geometry)

  ⟶ save the setup time
- Simple design (loop in cnt2)
- Simple to use (test case available – ask P. Marguinaud)
- Savings ~ 20 %

FA - LFI

- FA and LFI have a thread safe interface

  ⟶ possible to read/write multiple files using OpenMP
  - OK with GRIB Edition 0
  - Not yet with GRIB Edition 1 (aka Gribex)

IO server

- Field transposition is now performed by the IO server tasks (no more GATH_GRID/GATH_SPEC in the model)
- Synchronization tools (retrieve model data as soon as produced) ; test case available – ask P. Marguinaud
- The IO server creates several files for each forecast term

  → make the LFI library handle this transparently, no need to concatenate files (an index is created over multiple LFI files)
- Read coupling files with the IO server ; saves 2.5 % of time on AROME 1.3 km

Summary / Conclusion

- ***Ready for higher resolutions :***
  - Arome 1.3 km L90
  - Arpege T1198L105
- ***Among next plans :***
  - Toward higher scalability
    - Fullpos-2 in a single 4DVar binary for OOPS
    - ...
  - Continuing porting
    - Parallelization of FESTAT-Arpege
    - Conf. 903 including the « PP-server » facility to replace the monotask conf. 901 (« IFS-to-FA » file transformation)
    - ...

Thank you for your attention !