Technical Report

Porting of ALADIN Algeria research scripts on the Météo-France BULL computer (beaufix)

Realized by: DJEBBAR ARAB.
Supervisors: ALEXANDRE MARY & CLAUDE FISCHER.

Météo-France/Toulouse
October 14 to 26, 2013

Acknowledgements:

I wish to thank Rezzagui Ali who gave me the opportunity to come to Météo-France. Special thanks to Claude Fischer and Alexandre Mary, for their availability and much appreciated help during this visit.
I also thank all those who have contributed to make this stay a success on all fronts, including:
Mr Eric Escalière
Ms Taillefer Françoise
Mr Jean Maziejewski
Mr Pascal Raynaud

This stay was supported by the French Embassy in Algiers, within the bilateral MF/ONM cooperation program.

Table of contents

I Introduction:

As part of the bilateral cooperation between the national office of meteorology (ONM) and Météo-France (CNRM/GMAP for NWP), I did a stay of two weeks under the direction of Mr. CLAUDE FISCHER.
The purpose of this visit was to port the ALADIN-Algeria model research scripts (e923, e927, e001 and fullpos) on the new computer BULL of Météo-France.
To facilitate the work, the task was split into three parts which are, in the order:

I -Upgrading the Aladin model scripts from cycle 36 to cycle 38 on the current machine NEC at Météo-France.
II - Porting these scripts on Bull, by carrying out all the major changes.
III- Comparison of forecast outputs on both machines.

But first we will present a short description of the new Bull computer.

II Short description of new Bull Super-computer BEAUFIX:

1. Some Bull machine features:

The name of the machine Bull is 'beaufix'. This cluster will become by February 2014 MF's operational machine. Once the 2nd cluster (named “prolix”) will be installed in 2014, “beaufix” will remain the research cluster, and “prolix” will become the operational cluster.

- beaufix consists of 1080 compute nodes identified as 'beaufix [0-1079]'

- beaufix [0-71] , or 72 nodes with 128 Gigabytes (Gb) of memory.
- beaufix [72-79], or 8 nodes with 256 Gigabytes of memory.
- beaufix [80-1079] or 1000 nodes with 32 Gigabytes of memory.

Each node consists of 2 sockets (INTEL processor E5-2697 v2) and on each of them there are 12 processing cores (hearts) numbered from 0 to 11. Hyperthreading (HT) is enabled on beaufix, on each heart 2 compute threads (numbered 0 and 1) can be simultaneously active.
The command ***sinfo*** allows to know the availability and distribution of nodes.

By their memory size, nodes are grouped into different partitions (the equivalent of a queue under NQS):

- **ft-oper** : 2 nodes dedicated to transfers to cougar (reserved for operational)
- **oper** : 200 nodes 32Gb/node (reserved for operational).
- **normal32** : 350 nodes 32Gb/node (64 nodes and 10h elapse Max. /Job)
- **plenty32** : 378 nodes 32Gb/node (378 nodes and 6h elapse Max. /Job)
- **maxi** : 72 nodes 128Gb/node 728+nodes 32Gb/node a total of 780 nodes and 6 h elapse Max. /job)
- **large128** : 72 nodes 128 Gb/node (no time limit)
- **huge256**: 8 nodes 256 Gb/node (no time limit)
- **bonus32** : 72 nodes 32 Gb/node (no time limit)

state of a node:
-idle : unoccupied, available for work,
-alloc : busy, unreachable,

-drain : excluded for a potential problem (hardware or software).
-mixte : the processor is reserved only partially (all hearts
is not restricted.

Under beaufix the jobs are submitted via the utility SLURM (Simple Linux Utility for Resource Management).
It is through SLURM that jobs are submitted in batch mode.

2. Basic Commands:

SLURM provides several ways to submit a job and basically three commands are essential: sbatch, salloc and srun.

➔ sinfo... :
nodes availability.
sbatch job:
submission of commands in batch job, equivalent to qsub for submit on the NEC machines.
salloc ...:
allows you to allocate a set of nodes in order to perform commands later (for example, those in a separate job, especially srun)
srun ...:
will launch programs on allocated nodes (sequential or parallel with OpenMP or MPI or via bullxmpi IntelMPI)

sbatch = salloc + srun (so no salloc in the script launched by sbatch!)

III Upgrading the Aladin model cy36 to cy38 on the NEC machine:

As first step in updating cycle 36 to cycle 38, we have modified the input files(climatology files) in the coupling configuration(e927) as illustrated in this piece of e927 script code:

```
 C L I M
#---------- ------------------------------------------------

\cp
/mf/dp/marp/marp001/tampon/const/clim_arpege/tl798/clim_arpeg
e.tl798.01.m${MM} ${ENTREE}/clim_arpege.tl798.01.m${MM}

\cp  /cnrm/gp/mrpe/djebbara/clim_algerie.12km00.03.m${MM}
${ENTREE}/clim_algerie.m${MM}
-----------------------------------------------------------------
```

So we have modified the clim_arpege and clim_algerie files which deal with the new Aladin cy38.

Secondly, we have also changed the climatology files that are an input of the e001 configuration.
Because of the **in-line** Full-POS mode in the Aladin model cy38, the climatology inputs are done in both e001 and fullpos configurations. See below:

```
 #---------------coping clim-files-----------------------
--------

\cp -b 32768 ${HOME}/clim_algerie.12km00.03.m${MM} Const.Clim

\cp -b 32768 ${HOME}/clim_dap.alge01.03.m${MM} const.clim.ALGE01

#-----------------------------------------------------------------
```

clim_algerie,12km00.03 ---------->>> model clim

and

clim_dap.alge01.03--------------->>>latlon clim

Furthermore, the executable file (MASTER) and its namelist was changed to new

ones:
-cy38t1_master-op1.07.SX20r441.x.exe
-al38t1_algerie-op1.02.nam

For validation, at the end of this first phase, we have made several RUNS of the new Aladin cy38 on NEC.

IV Porting of the Aladin research scripts on beaufix machine:

Before any practical work can start, one needs to define his/her work environment on beaufix.
For instance, in my own case, I first connected by ssh as: djebbara@beaufix into the bull machine. Then, I introduced my login and password.
Once connected and cd/moved to my working directory, the prompt looks as below:

```
djebbara@beaufixlogin0: ll
e923
e927
e001
fulpos
```

Adapting the script from yuki to beaufix required major changes in the different configurations of the Aladin cy38.
Firstly, I had to rewrite the features cards (#SBATCH...) in the job's header, in all configurations(namely e923, e927, e001 and fullpos).
The differences in the header, between NEC and bull versions, are listed in the following Table:
example of e927 features card:

Beaufix card	NEC card
#!/bin/bash #SBATCH -p normal32 #SBATCH -n 24 #SBATCH -c 1 #SBATCH -N 1 #SBATCH -t 00:30:00	#!/bin/bash #PBS -S /bin/ksh #PBS -N e927 #PBS -j o #PBS -T mpisx #PBS -l elapstim_req=01:30:00

#SBATCH --job-name="e927" #SBATCH --mem=30000 #SBATCH --exclusive #SBATCH -verbose # -N : nombre de noeuds # -n : nombre TOTAL de taches MPI # -c : nombre de threads (OMP) # -t: elapse time # -p: nods class # -mem: memory	#PBS -l cputim_job=01:20:00 #PBS -l cpunum_job=2 #PBS -b 1 #PBS -l memsz_job=26000mb #PBS -q vector

All environment variables have been adapted to the characteristics of the beaufix calculator, as follows:

```
-------------------------------------------------------------
# Job management :
# --------------
export AUTO_PROFILE_HT="on"

# Number of nodes/mpi-tasks/omp-threads:
# -------------------------------------
NNODES=$SLURM_JOB_NUM_NODES
# Number of MPI tasks per node:
MPITASKS_PER_NODE=$((SLURM_NTASKS/SLURM_JOB_NUM_NODES))
# Number of OPEN-MP threads per MPI task:
export OMP_NUM_THREADS=$SLURM_CPUS_PER_TASK
# Total number of MPI tasks:
MPI_TASKS=$SLURM_NTASKS
NPROC=$MPI_TASKS

echo NNODES=$NNODES
echo MPITASKS_PER_NODE=$MPITASKS_PER_NODE
echo MPI_TASKS=$MPI_TASKS
echo OMP_NUM_THREADS=$OMP_NUM_THREADS

# Specific environment variables :
# ------------------------------
set -x
#export OMP_STACKSIZE=1G
export KMP_STACKSIZE=1G
export KMP_MONITOR_STACKSIZE=1G
set +x
ulimit -s unlimited

# Software default environment variables :
```

```
# ------------------------------------
set -x
export DR_HOOK=1
export DR_HOOK_IGNORE_SIGNALS=-1
export DR_HOOK_OPT=prof
export MPL_MBX_SIZE=2048000000
export EC_PROFILE_HEAP=0
export EC_MPI_ATEXIT=0
export MPIAUTOCONFIG="mpiauto.TIME.conf"
#------------------------------------------------------------
```

After the header and environment variables Sections, comes the script section where clim files and binary (executable) files are copied to the local directory (Note: those files already are ported on beaufix before), as shown below:

- ✔ e927 configuration

```
#---------------* clim files coping *-----------------------

cp ${HOME}/clim_algerie.12km00.05.m${MM} Const.Clim

cp ${HOME}/clim_dap.alge01.05.m${MM} const.clim.ALGE01
 const.clim.ALGE01
```

- ✔ e001 configuration

```
#-----------------*Executable*------------------------------ -
cp ${HOME}/cy38t1_master-op2.10.IMPI411IFC1301.2x.exe MASTER
#------------------------------------------------------------ --
```

- ✔ fullpos configuration

```
-----------------* progrid *-------------------------------------
cp ${HOME}/ut38t1_progrid-op2.09.IMPI411IFC1301.2x.exe PROGRID
  ----------------------------------------------------------------------------------------------------
```

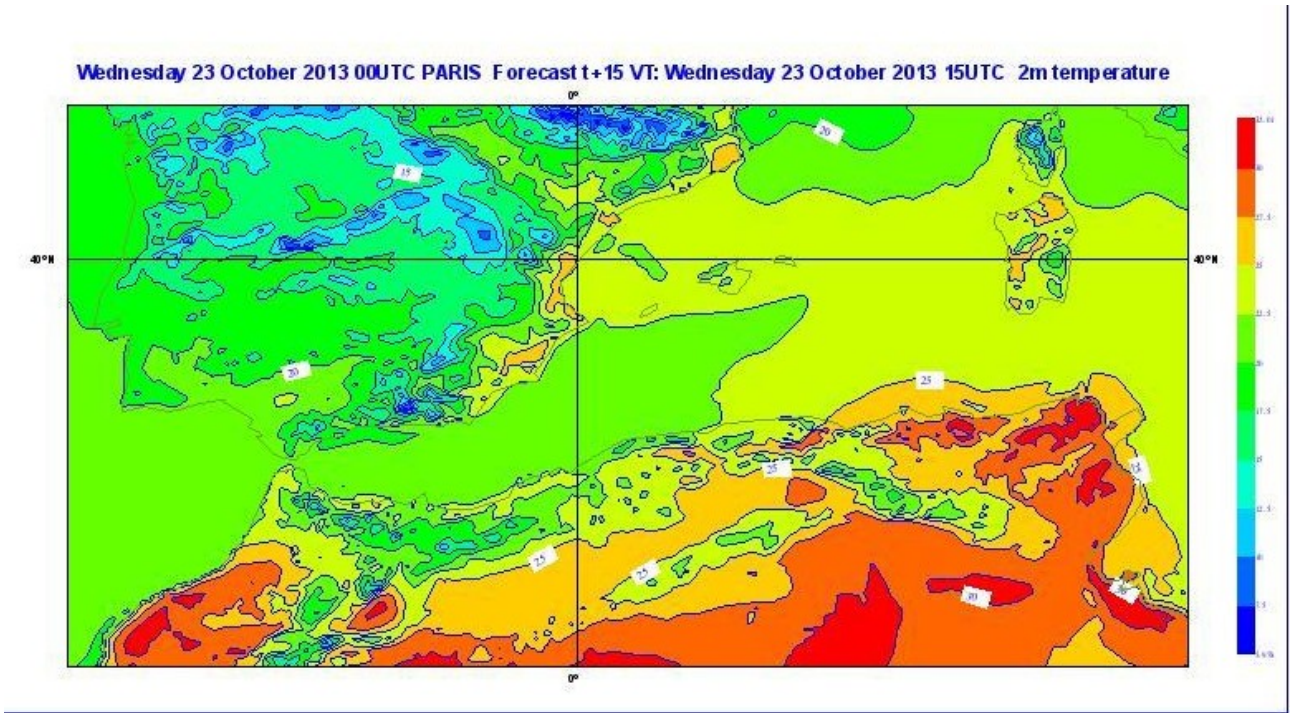
- ➔ The commands used to execute the Master (binary files) in all our scripts are as follows:

```
MPIAUTO=/home/gmap/mrpm/marguina/SAVE/mpiauto/mpiauto

# MPI job :

time $MPIAUTO -np $MPI_TASKS -nnp $MPITASKS_PER_NODE --wrap
--verbose -- ./$EXECUTABLE $ARGUMENTS </dev/null
```

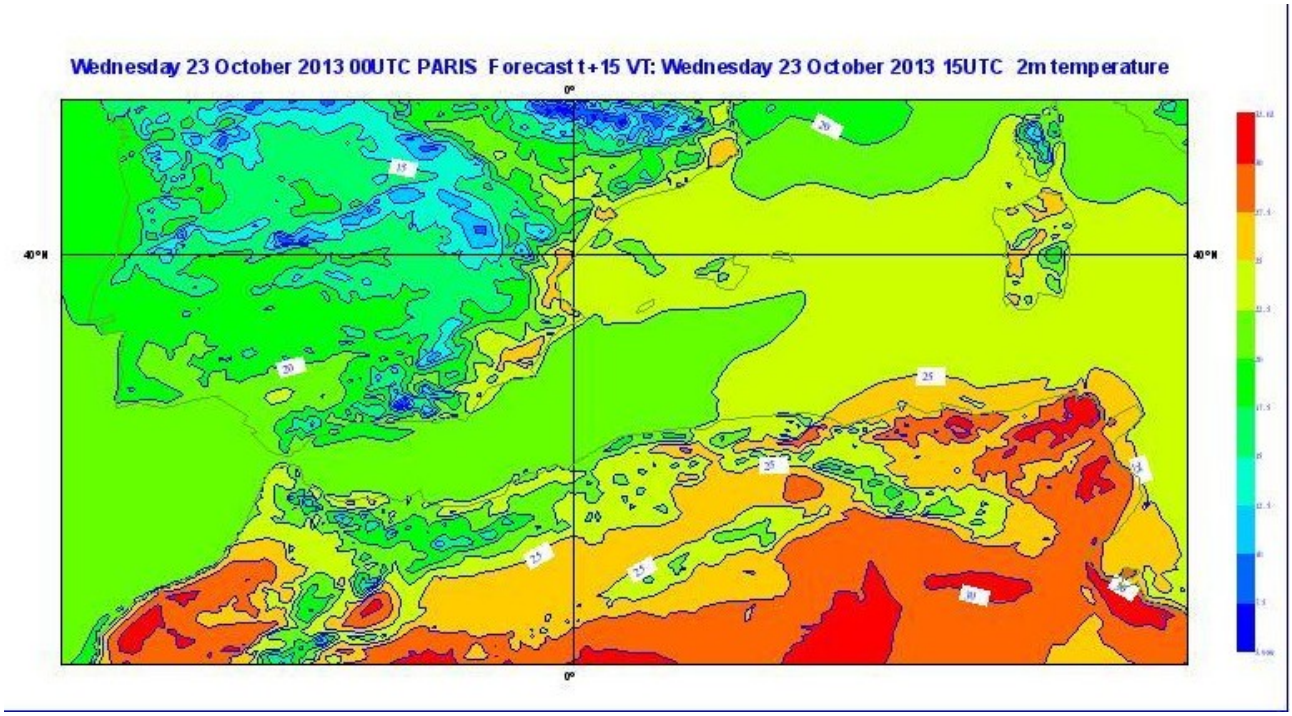
V Comparison of some Aladin forecast fields produced on NEC and on beaufix machines:

1- Temperature at t + 15h:

a) On Beaufix

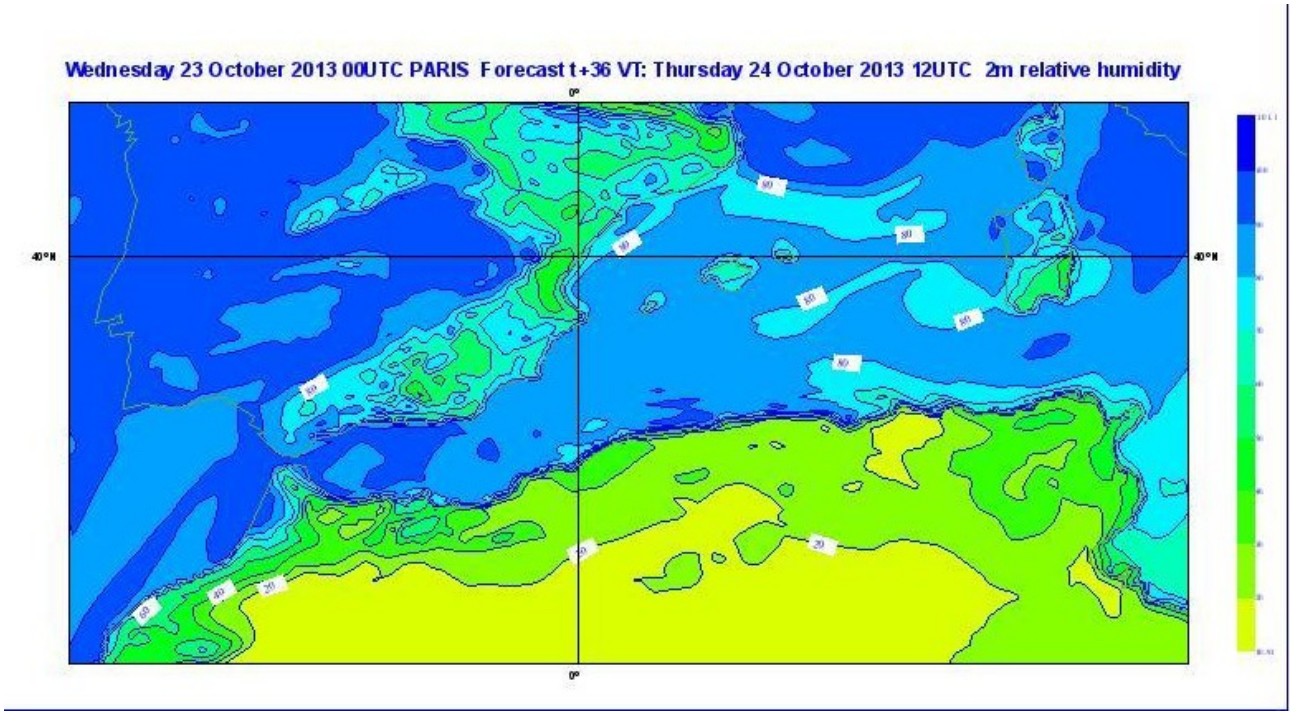


b) On NEC

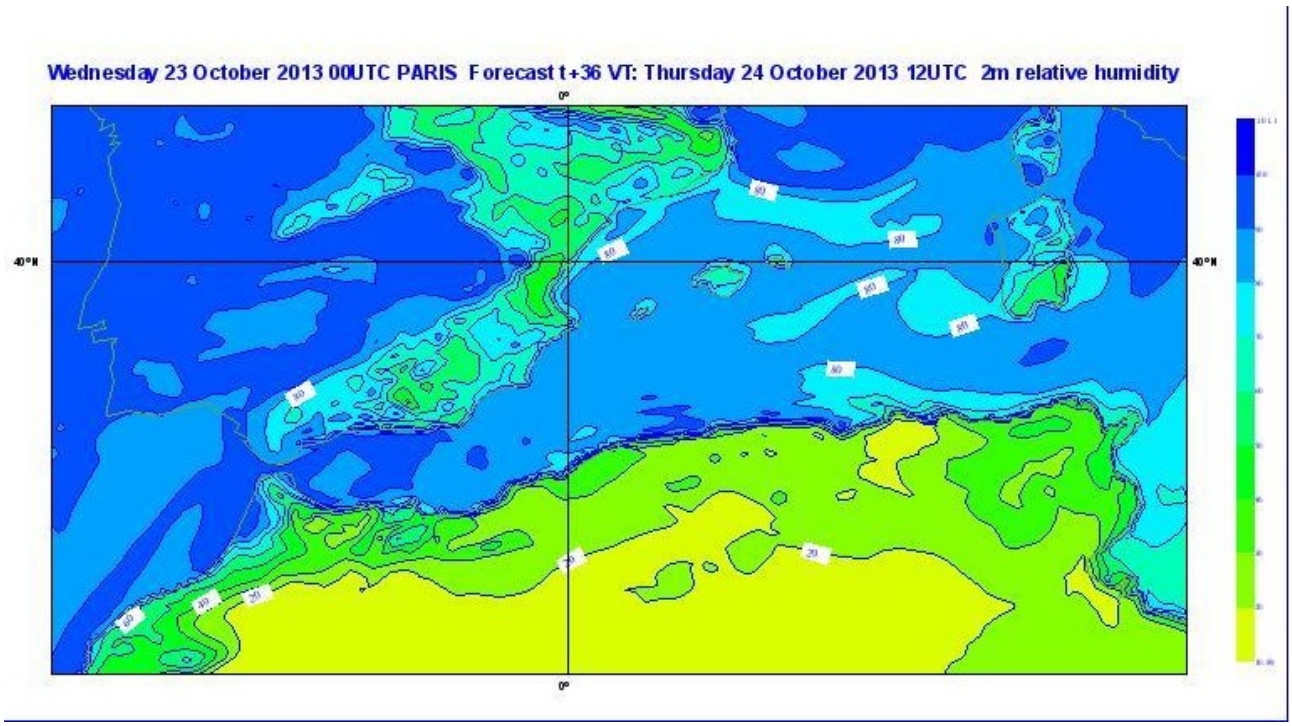


2- Humidity at t +36h

a) On Beaufix

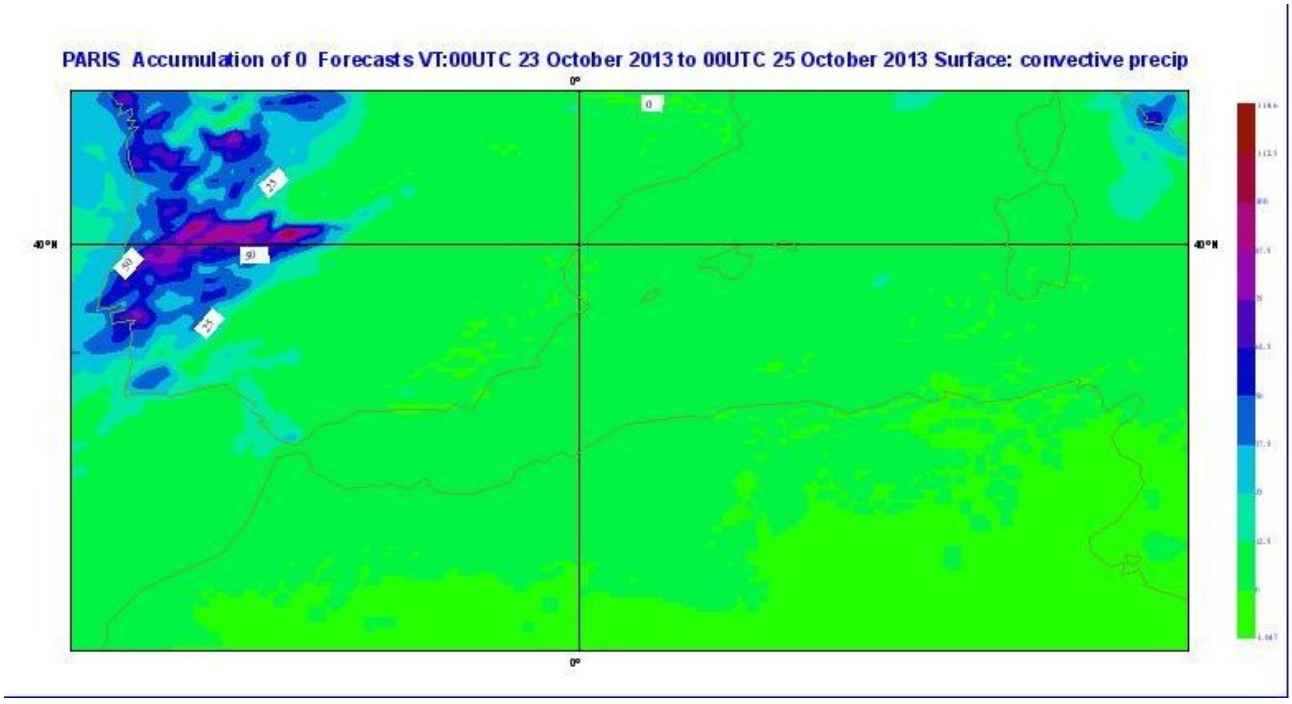


b) On NEC

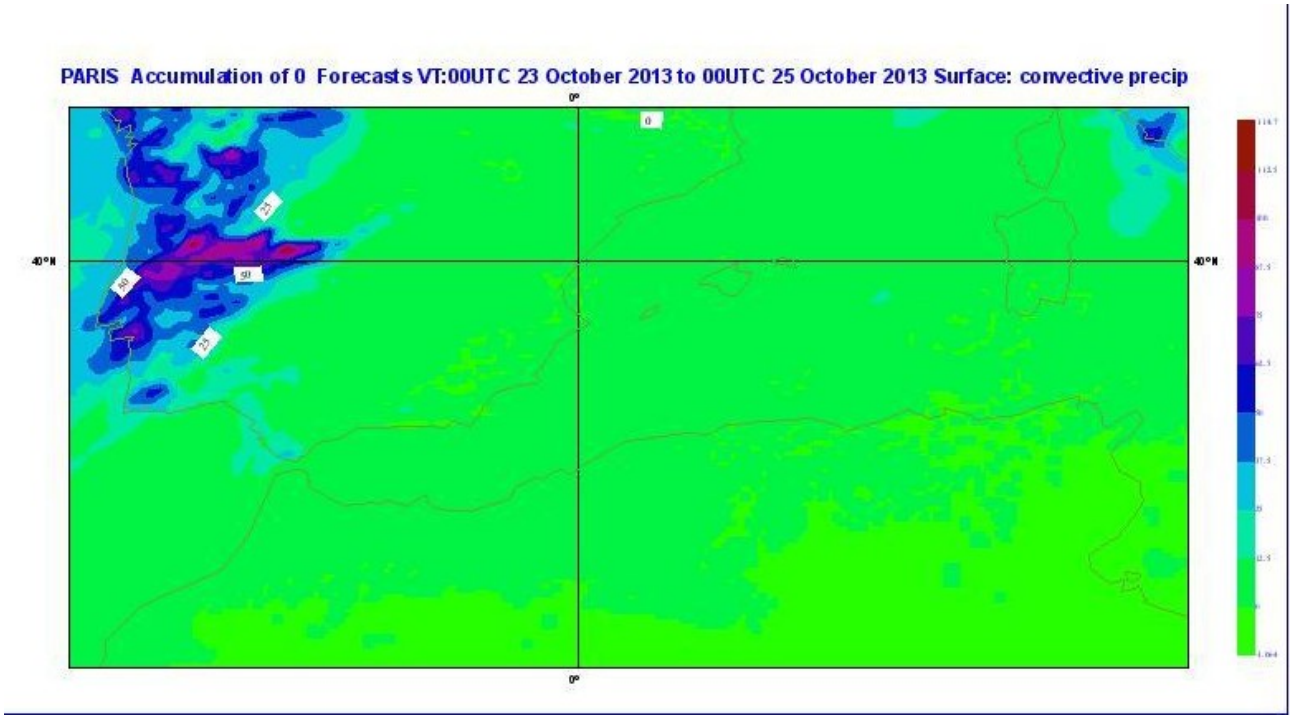


3-Total rain : 0 to 48h

a) On Beaufix



b) On NEC



As can be seen in these forecasts, the outputs of the two models are similar

As a general conclusion for this work, a set of new scripts valid for MF's new super-computer beaufix has been prepared (e923, e927, e001, Full-POS). These scripts can be made available for adaptation for other Aladin partners who'd wish to run these configurations on beaufix from home, for their own R&D activity. For the case of Aladin-Algeria, some comparisons have been made in order to assess the meterological neutrality of the change from NEC/yuki to BULL/beaufix. The comparisons provided satisfactory results.