pyGEFS: PYTHON based workflow management of NCEP Global Ensemble Forecast System

Xianwu Xue[1]

Dingchen Hou[2], Walter Kolczynski[3], Yuejian Zhu[2], Bing Fu[3], Xiaqiong Zhou[4], Eric Sinsky[3], Wei Li[3], Hong Guan[1] and Bo Cui[3]

[1]SRG at Environmental Modeling Center, NCEP/NWS, College Park, MD
[2]Environmental Modeling Center, NCEP/NWS, College Park, MD
[3]IMSG at Environmental Modeling Center, NCEP/NWS, College Park, MD
[4]NOAA/GFDL, Princeton University Forrestal Campus, Princeton, NJ

Acknowledgments

EMC ensemble team, Jack Kain, Wen Meng, Richard Wobus and Terry McGuinness

100th AMS Annual Meeting - January 12-16, 2020, Boston, MA

Challenges in the development of GEFS v12

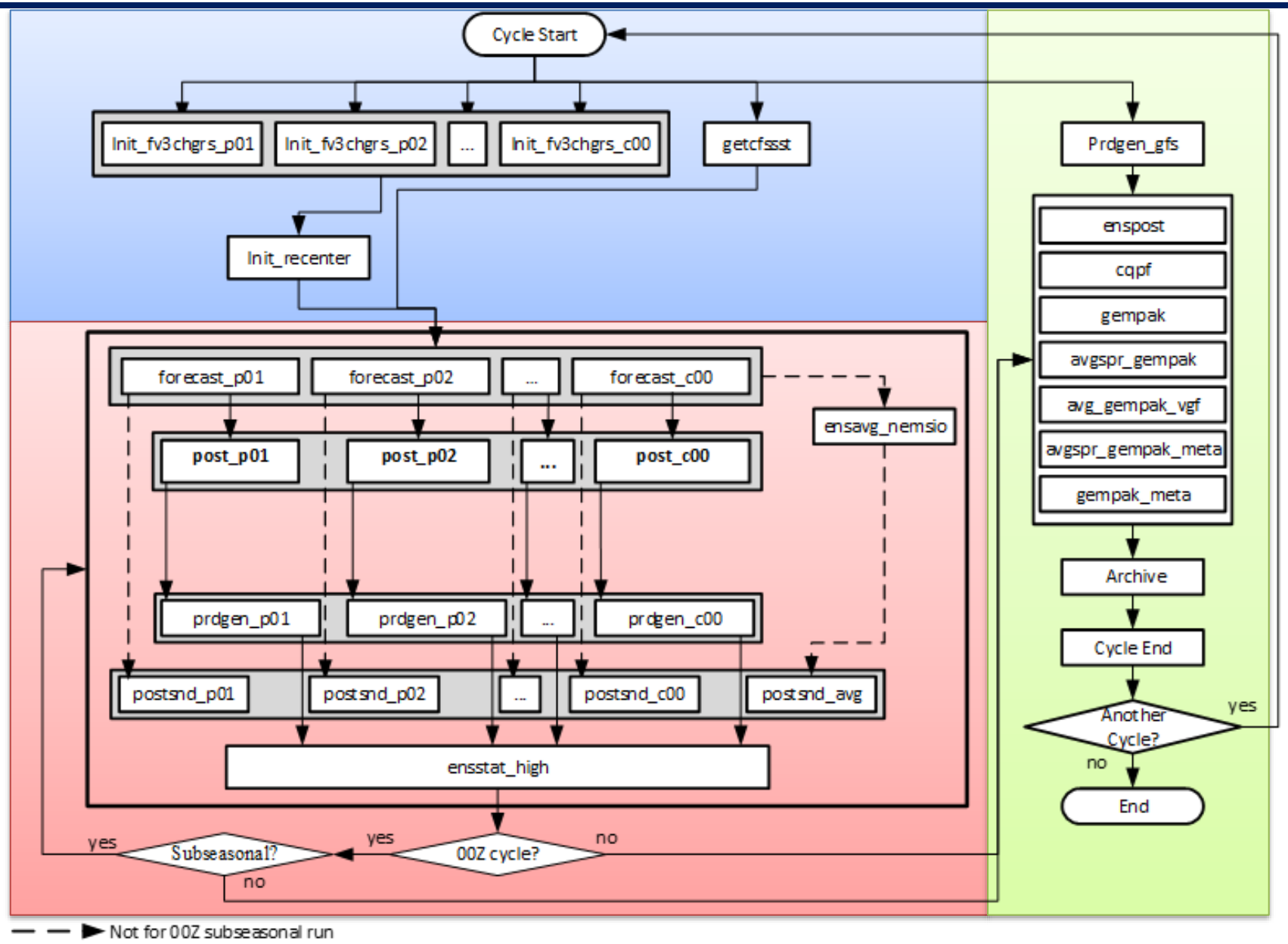
- Science and Scope
  - Adding new capabilities
  - Coupling with wave model (One Way)
  - Coupling with aerosol model (One Way)
  - Extending scope of products
- Implementation Requirements
  - Reforecast – Phase 1 and 2 (30 Years)
  - Retrospective Atmosphere Only (2.5 Years)
  - Retrospective Atmosphere + Wave (1 Year)
  - Retrospective Atmosphere + Aerosol (1 Year)
  - Producing protype ecFlow scripts
- Working on the Different Platforms
  - NCEP operational machines (WCOSS P1/2, Cray, DELL 3, DELL 3.5 (Future) …)
  - NOAA research machines (Theia, Hera, Jet …)
  - Cloud (AWS)

Simplifying and optimizing workflow management is one of the critical tasks in the development and implementation of GEFS v12

Flowchart of GEFS – Atmosphere

Initialization

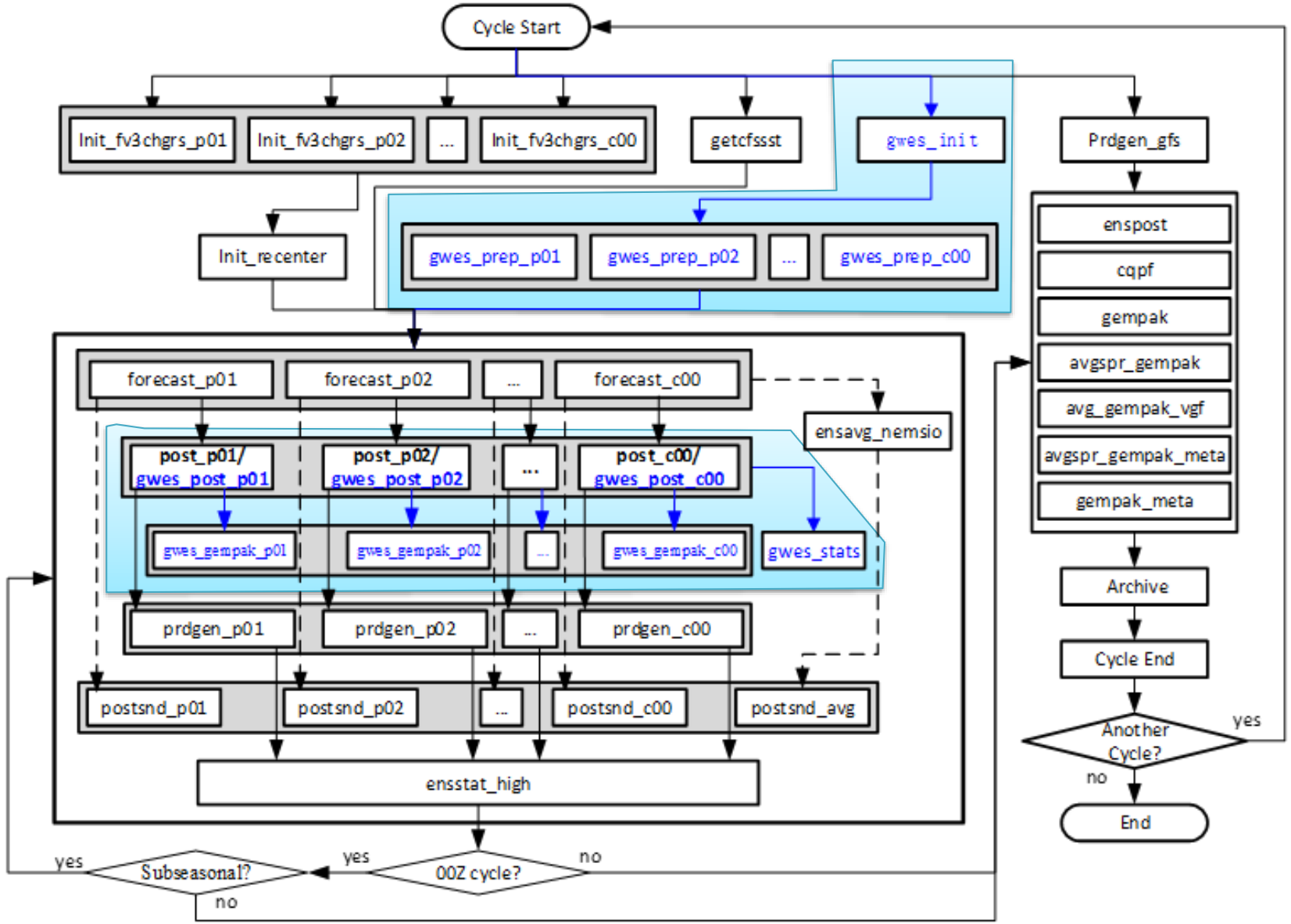
Forecast and Post



Products and Data Archive

Flowchart of GEFS – Atmosphere + Wave

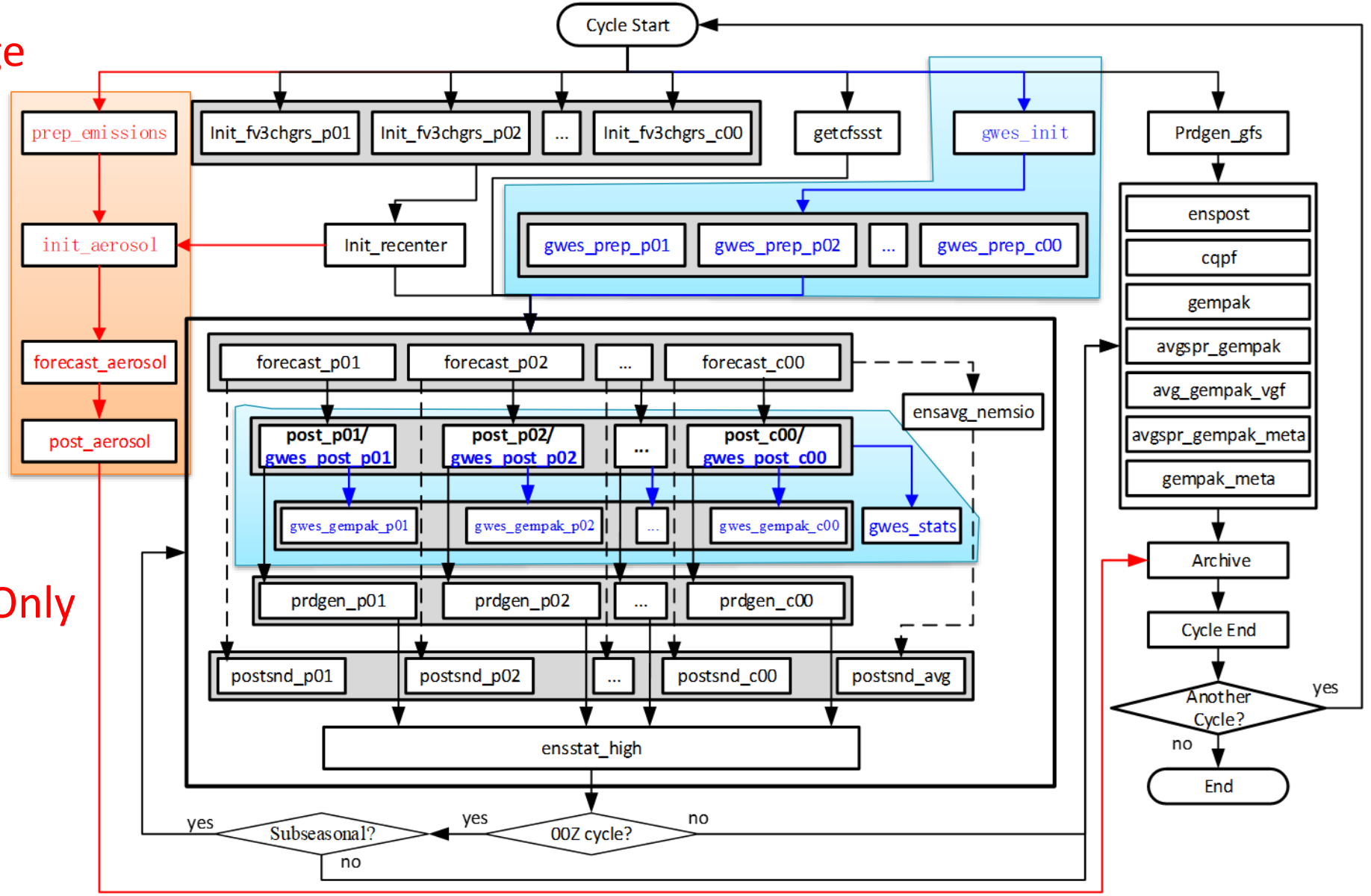
More Complicated



Flowchart of GEFS – Atmosphere + Wave + Aerosol

Final Package

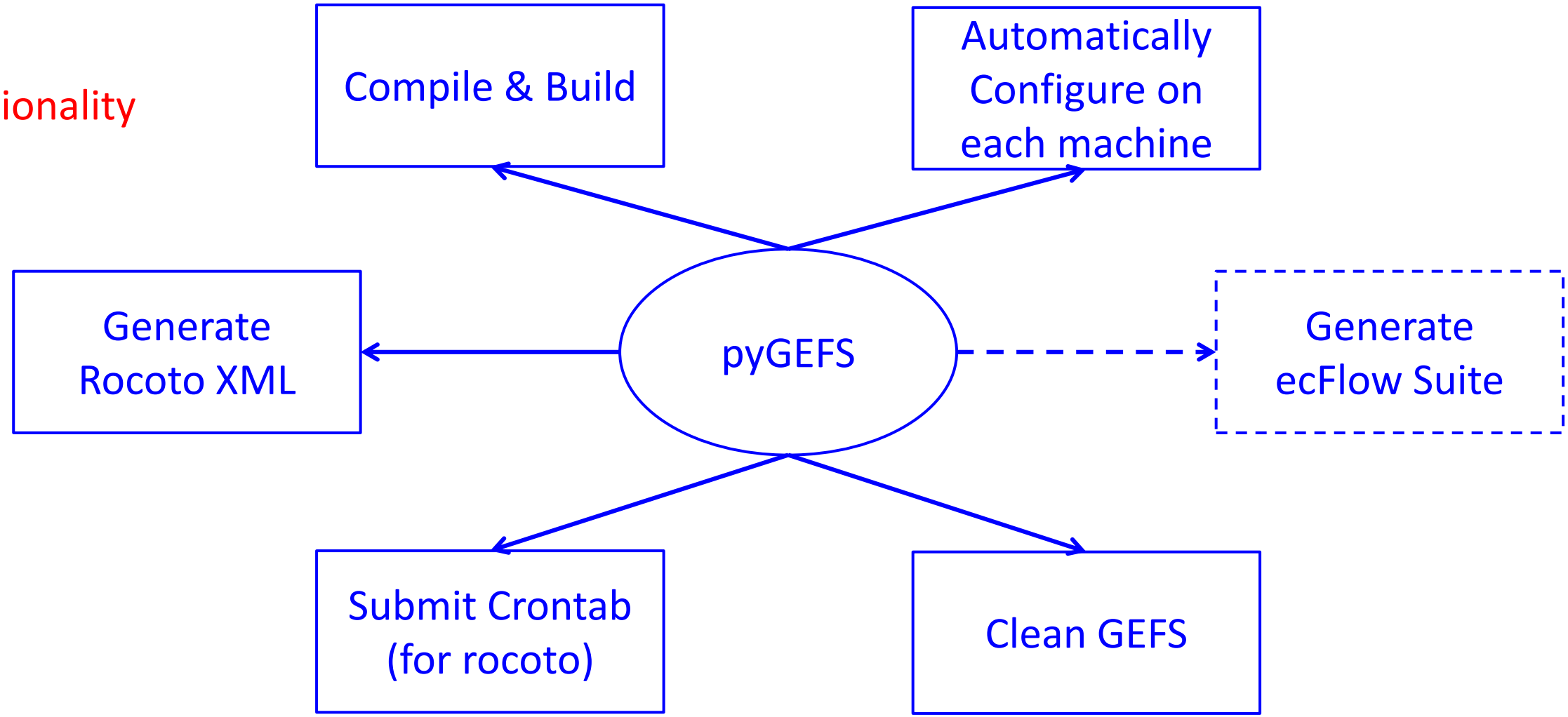
Aerosol Control Only



What is pyGEFS?

pyGEFS: A group of python scripts, supporting shell scripts and test files used to manage the workflow of GEFS

Functionality



--- → Working on

Legacy Configuration and run GEFS

- clone GEFS
- cd GEFS/sorc
- mkdir ../exec
- mkdir ../util/exec
- module purge
- module use ./
- module load Module_gefs_v12_machine [machine: cray, wcoss, hera …]
- ./build.sh
- ./install.sh
- cd ../rocoto
- …

Configure and run GEFS using pyGEFS

- clone GEFS
- cd GEFS/rocoto
- ./compile_install_all.sh -c yes -r yes -b yes [-f user_full.conf]

Compile and Build

Generate workflow

Add to crontab

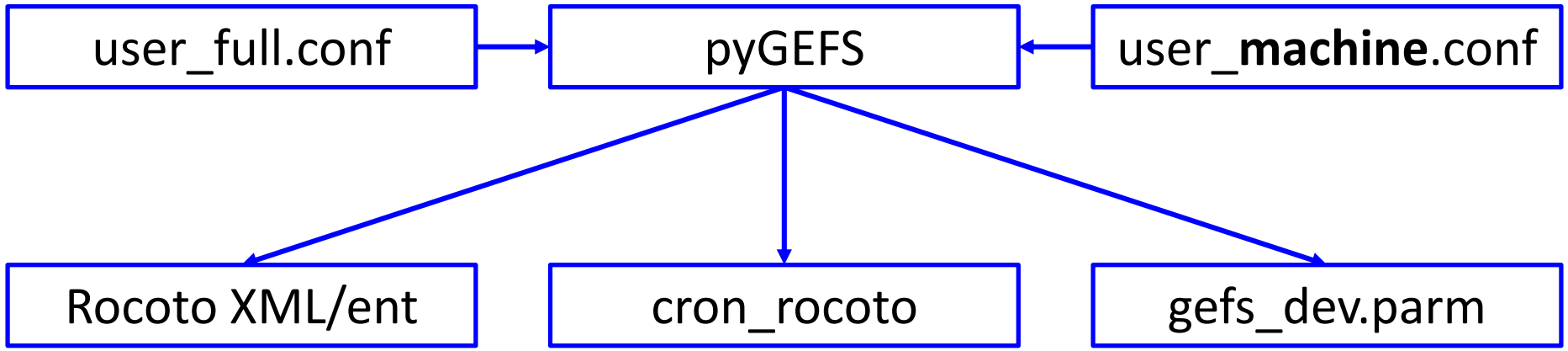
Optional Specific Configure File

```
 CYCLE                         TASK                                JOBID      STATE         EXIT TRIES DURATION
=== (updated:2019-12-17 21:39) ================== PSLOT: XX_Master_NewBuild ======================================
201912140000    getcfssst                                         13702096    SUCCEEDED        0     1    333
201912140000    gwes_init                                         13702097    SUCCEEDED        0     1    32
201912140000 <  gwes_prep                                         13702201    31/31 SUCCEEDED  0     1    10
201912140000 <  init_fv3chgrs                                     13702098    31/31 SUCCEEDED  0     1    157
201912140000    init_recenter                                     13703390    SUCCEEDED        0     1    336
201912140000 <  forecast_high                                     13703955    31/31 SUCCEEDED  0     1    7047
201912140000 <  post_high                                         13712465    31/31 SUCCEEDED  0     1    3904
201912140000 <  prdgen_high                                       13712736    31/31 SUCCEEDED  0     1    3219
201912140000    ensstat_high                                      13714947    SUCCEEDED        0     1    3663
201912140000 <  gwes_post                                         13712502    31/31 SUCCEEDED  0     1    1170
201912140000    prdgen_gfs                                        13703007    SUCCEEDED        0     1    8872
201912140000 <  forecast_low                                      13712517    31/31 SUCCEEDED  0     1    7967
201912140000 <  post_low                                          13715162    31/31 SUCCEEDED  0     1    4977
201912140000 <  prdgen_low                                        13715242    31/31 SUCCEEDED  0     1    4689
201912140000    ensstat_low                                       13719993    SUCCEEDED        0     1    1623
201912140000    gempak                                            13719848    SUCCEEDED        0     1    1191
201912140000    avgspr_gempak                                     13719849    SUCCEEDED        0     1    1188
201912140000    avg_gempak_vgf                                    13720594    SUCCEEDED        0     1    11
201912140000    avgspr_gempak_meta                                13720601    SUCCEEDED        0     1    38
201912140000    gempak_meta                                       13720602    SUCCEEDED        0     1    239
201912140000    ensavg_nemsio                                     13715167    SUCCEEDED        0     1    3101
201912140000 <  postsnd                                           13715168    32/32 SUCCEEDED  0     1    690
201912140000 <  post                                              13719850    2/2 SUCCEEDED    0     1    13
201912140000    enspost                                           13720736    SUCCEEDED        0     1    2284
201912140000 <  keep_data                                         13721763    2/2 SUCCEEDED    0     1    2648
201912140000 <  archive                                           13721764    2/2 SUCCEEDED    0     1    3449
201912140000    cleanup_atm                                       13722970    SUCCEEDED        0     1    224
201912140000    cleanup_wave                                      13720603    SUCCEEDED        0     1    9
```

RocotoViewer

Input of pyGEFS

- user configuration file
  - user_full.conf
    - py/user_wcoss_dell_p3.conf
    - py/user_cray.conf
    - py/user_hera.conf



./compile_install_all.sh -r yes

1) Flexibility of the Configure File

```
#SOURCEDIR              = /gpfs/dell2/emc/retros/noscrub/Dingchen.Hou/TAGS/DHrestart
#
SDATE                   = 2019110500 # Start Date and Time
EDATE                   = 2019110500
npert                   = 30         # Ensemble Size
INCYC                   = 24
#ACCOUNT                 = GEN-T2O
#CUE2RUN                 = dev
#TRANSFER_QUEUE          = dev_transfer
#SCHEDULER               = lsf
#HPS_PTMP                = dell2
CYCLE_THROTTLE          = 1
TASK_THROTTLE           = 65
# Start Parm ++++++++++++++++++++++++++ For gefs_dev.parm, ++++++++++++++++++++++
#
# Define data streams for prdgen. Each stream will run in a separate thread.
# Streams res_2p50, res_1p00, res_0p50, res_0p25_s1, res_0p25_s2, and res_0p25_s3 are
# defined, but other streams can be defined here. Instructions are in parm/gefs_prdgen
# If you change the number of streams, be sure to update the CPU request for gefs_prdg
# gefs_prdgen_gfs below, and also modify the rocoto execution script (in rocoto/bin/<m
# correspondingly.
#
# PRDGEN_STREAMS    =   "res_2p50 res_1p00 res_0p50 res_0p25_s1 res_0p25_s2 res_0p25_s
#
# If you want to define a new stream, you must define the following for each new strea
#   PRDGEN_GRID[stream_name]="2p5"                          # old jobgrid, used to
#   PRDGEN_GRID_SPEC[stream_name]="latlon 0:144:2.5 90:73:-2.5" # grid specification f
#   PRDGEN_HOURS[stream_name]="{0..384..6}"                # forecast hours to cr
#   PRDGEN_SUBMC[stream_name]="prd2p5"                     # temporary directory
#   PRDGEN_A_DIR[stream_name]="pgrb2a2p5"                  # directory name for p
#   PRDGEN_B_DIR[stream_name]="pgrb2b2p5"                  # directory name for p
#   PRDGEN_A_PREFIX[stream_name]="pgrb2a.2p50."            # pgrba identifier in
#   PRDGEN_B_PREFIX[stream_name]="pgrb2b.2p50."            # pgrbb identifier in
#
# Setting these are optional for a new stream (default=NO)
#   PRDGEN_DO_ANALYSIS[stream_name]="NO"
#
# For subjobs
N_SUBJOBS_POST_HIGH     = 0
N_SUBJOBS_ENSAVG_NEMSIO = 0
GEMPAK_RES              = "1p00 0p50" #1p00 0p50 0p25
save_pgrb2_p5           = NO
save_pgrb2_p25          = NO
fhmax                   = 384
fhmaxh                  = 384
FHMAXHF                 = 240
```

Separate the workflow management from the main GEFS

SOURCEDIR = **PathToGEFS**

```
SDATE                   = 2019110500 # Start Date
EDATE                   = 2019110500
npert                   = 30         # Ensemble Size
```

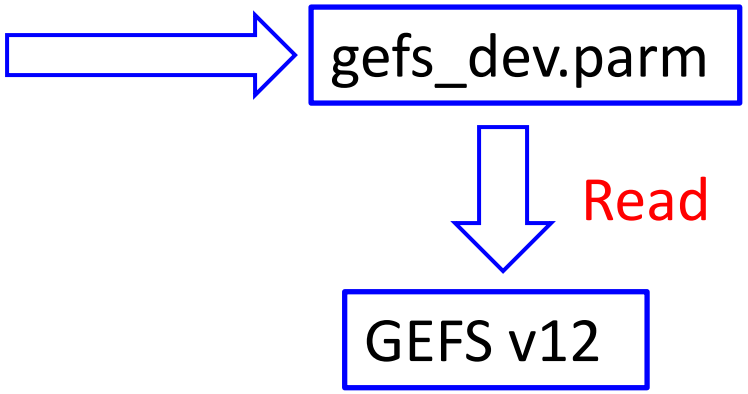
Add description to the variable

2) Flexibility to add/modify parameters passed to GEFS

```
# Start Parm ++++++++++++++++++++++++++++ For gefs_dev.parm, ++++++++++++++++++++++++++++\/\/
# For subjobs
N_SUBJOBS_POST_HIGH     = 0
N_SUBJOBS_ENSAVG_NEMSIO = 0
GEMPAK_RES              = "1p00 0p50" #1p00 0p50 0p25
save_pgrb2_p5           = NO
save_pgrb2_p25          = NO
fhmax                   = 384
fhmaxh                  = 384
FHMAXHF                 = 240
FHOUTHF                 = 3
FHOUTLF                 = 6
VERBOSE                 = no
# navg_min < npert
navg_min                = 10
#define tmp time step
DELTIM                  = 450
k_split                 = 2
n_split                 = 6
TYPE                    = nh
MONO                    = non-mono
# cpu geometry
#layout_x                = 8
#layout_y                = 8
#WRITE_GROUP             = 1
#WRTTASK_PER_GROUP       = 64
#parallel_threads        = 1
# others
ENS_SPS                 = .false.
DO_SPPT                 = YES
DO_SHUM                 = NO
DO_SKEB                 = YES
#RFHOME                  = /gpfs/dell3/nco/storage/fv3gefs
#enslistend              = "avg spr"
#SSTDIR                  = $HOMEdata/2tsst
# End Parm ++++++++++++++++++++++++++++ For gefs_dev.parm, ++++++++++++++++++++++++++++/\/\
```

Modify/add variables (as below) for both environmental and scientific parameters

```
export var=${var:-0}
```



3) Flexibility to run different task configurations

```
RUN_GETCFSSST           = YES
RUN_WAVE_PREP           = YES
RUN_INIT                = FV3_COLD
KEEP_INIT               = NO
RUN_FORECAST_HIGH       = YES
RUN_POSTSND             = NO
RUN_PRDGEN_GFS          = NO
RUN_FORECAST_LOW        = NO
RUN_GEMPAK              = NO
RUN_TRACK               = NO
RUN_OTHERS              = NO
RUN_CLEANUP             = NO
##### taskname is for development, if you have taskname, RUN_* will NOT be used
#taskname               = gwes_init
#taskname               = gwes_prep
#taskname               = rf_prep
#taskname               = getcfssst
#taskname               = prdgen_gfs
#taskname               = init_fv3chgrs
#taskname               = init_recenter
#taskname               = keep_init
#taskname               = copy_init
#taskname               = forecast_high
#taskname               = post_high
#taskname               = gwes_post
#taskname               = prdgen_high
#taskname               = ensstat_high
#taskname               = gwes_stats
#taskname               = ensavg_nemsio
#taskname               = postsnd
#taskname               = forecast_low
#taskname               = post_low
#taskname               = prdgen_low
#taskname               = ensstat_low
#taskname               = gempak
#taskname               = avgspr_gempak
#taskname               = avgspr_gempak_meta
#taskname               = avg_gempak_vgf
#taskname               = gempak_meta
#taskname               = extractvars
#taskname               = post_track
#taskname               = post_genesis
#taskname               = enspost
#taskname               = keep_data_atm
#taskname               = keep_data_wave
#taskname               = archive_atm
#taskname               = archive_wave
#taskname               = cleanup_atm
#taskname               = cleanup_wave
```

Either to RUN selected group(s) of tasks

Or to RUN specific task(s)

Good for Development

4) Flexibility to modify resources of existing task and add new task

Existing Task in user_full.conf

```
## extractvars************************************
#extractvars_walltime     = 00:30:00
#extractvars_nodes        = 1
#extractvars_ppn          = 1
#extractvars_tpp          = 1
#extractvars_memory       = 3000M
#extractvars_queue        = &CUE2RUN;
#extractvars_dep          = <taskdep task="ensstat_high"/>
```

```
## extractvars************************************
#extractvars_walltime     = 00:30:00
#extractvars_nodes        = 1
#extractvars_ppn          = 1
#extractvars_tpp          = 1
extractvars_memory        = 6000M
#extractvars_queue        = &CUE2RUN;
#extractvars_dep          = <taskdep task="ensstat_high"/>
```

```
## new_task***************************************************
new_task _walltime     = 01:30:00
new_task _nodes        = 1
new_task _ppn          = 2
new_task _tpp          = 2
new_task _memory       = 5000M
new_task _join         = &LOG_DIR;/@Y@m@d/gefs_extractvars_@H.%J
new_task _queue        = &CUE2RUN;
new_task _dep          = <taskdep task="ensstat_high"/>
```

New Task

./compile_install_all.sh -r yes

5) Flexibility to modify the dependency of a task

1. Add dependency automatically
   a) user_**machine**.conf
   b) programmed in scripts
2. Manually modify user configure file

```
## extractvars**********************************************
extractvars_walltime     = 00:30:00
extractvars_nodes        = 1
extractvars_ppn          = 1
extractvars_tpp          = 1
extractvars_memory       = 3000M
extractvars_join         = &LOG_DIR;/@Y@m@d/gefs_extractvars_@H.%J
extractvars_queue        = &CUE2RUN;
extractvars_dep          = <taskdep task="ensstat_high"/>
```

Modify manually

```
extractvars_dep          =
```

Means no dependency for this task

Summary

- Main features of pyGEFS
  - High portability/standalone package
  - High flexibility and expandability
  - Easy to modify and adapt
  - Compatibility for upgrade (version)
- Application
  - Model Development on all NOAA machines
  - Running GEFS Reforecast
  - Running GEFS Retrospective Forecast
  - Generating ecFlow scripts

Real-Time Monitoring System for Reforecast

The Real-Time Monitor System for Reforecast

SUCCEEDED -- RUNNING -- DEAD or FAILED

Year	Month											
1989	01	02	03	04	05	06	07	08	09	10	11	12
1990	01	02	03	04	05	06	07	08	09	10	11	12
1991	01	02	03	04	05	06	07	08	09	10	11	12
1992	01	02	03	04	05	06	07	08	09	10	11	12
1993	01	02	03	04	05	06	07	08	09	10	11	12
1994	01	02	03	04	05	06	07	08	09	10	11	12
1995	01	02	03	04	05	06	07	08	09	10	11	12
1996	01	02	03	04	05	06	07	08	09	10	11	12
1997	01	02	03	04	05	06	07	08	09	10	11	12
1998	01	02	03	04	05	06	07	08	09	10	11	12
1999	01	02	03	04	05	06	07	08	09	10	11	12
2000	01	02	03	04	05	06	07	08	09	10	11	12
2001	01	02	03	04	05	06	07	08	09	10	11	12
2002	01	02	03	04	05	06	07	08	09	10	11	12
2003	01	02	03	04	05	06	07	08	09	10	11	12
2004	01	02	03	04	05	06	07	08	09	10	11	12
2005	01	02	03	04	05	06	07	08	09	10	11	12
2006	01	02	03	04	05	06	07	08	09	10	11	12
2007	01	02	03	04	05	06	07	08	09	10	11	12
2008	01	02	03	04	05	06	07	08	09	10	11	12
2009	01	02	03	04	05	06	07	08	09	10	11	12
2010	01	02	03	04	05	06	07	08	09	10	11	12
2011	01	02	03	04	05	06	07	08	09	10	11	12
2012	01	02	03	04	05	06	07	08	09	10	11	12
2013	01	02	03	04	05	06	07	08	09	10	11	12
2014	01	02	03	04	05	06	07	08	09	10	11	12
2015	01	02	03	04	05	06	07	08	09	10	11	12
2016	01	02	03	04	05	06	07	08	09	10	11	12
2017	01	02	03	04	05	06	07	08	09	10	11	12
2018	01	02	03	04	05	06	07	08	09	10	11	12
2019	01	02	03	04	05	06	07	08	09	10	11	12

2013 - The Real-Time Monitor System for Reforecast

SUCCEEDED -- RUNNING -- DEAD or FAILED

Monitoring reforecast quality through verification: Time Series of Scores

January | February | March

April | May | June

July | August | September

October | November | December

All Files for all members
All Files for 11 members
All Files for 5 members

All crontal data were generated

02/2013 - The Real-Time Monitor System for Reforecas

SUCCEEDED -- RUNNING -- DEAD or FAILED

February

Sun	Mon	Tue	Wed	Thu	Fri	Sat
					01	02
03	04	05	06	07	08	09
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

02/08/2013 - The Real-Time Monitor System for Reforecast

SUCCEEDED -- QUEUED -- RUNNING -- DEAD or FAILED

reforecast_19Y_test

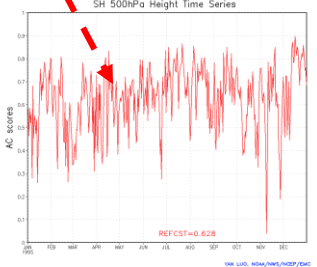
GEFS_ROCOTO: /gpfs/gd1/emc/ensemble/noscrub/emc.enspara/Hong.Guan/reforecast_19Y_test/rocoto

WORKDIR: /gpfs/gp2/ptmp/emc.enspara/Hong.Guan/o/reforecast_19Y_test

HPSS_DIR: /NCEPDEV/emc-ensemble/5year/emc.enspara/fv3gefs/REFCST

KEEP_DIR: /gpfs/dell3/nco/storage/fv3gefs/REFCST

CYCLE	TASK	JOBID	STATE	EXIT	TRIES	DURATION(m)	Start-Time	End-Time	DeltaT(m)
201302080000	rf_prep	4082612	SUCCEEDED	0	1	1	2019-09-24 03:15:08	2019-09-24 03:16:05	1
201302080000	forecast_low_p01	4084730	SUCCEEDED	0	1	144	2019-09-24 05:20:08	2019-09-24 07:44:15	144
201302080000	forecast_low_p02	4084731	SUCCEEDED	0	1	144	2019-09-24 05:20:08	2019-09-24 07:44:11	144
201302080000	forecast_low_p03	4084732	SUCCEEDED	0	1	144	2019-09-24 05:20:08	2019-09-24 07:43:56	144
201302080000	forecast_low_p04	4084733	SUCCEEDED	0	1	144	2019-09-24 05:20:08	2019-09-24 07:44:11	144
201302080000	forecast_low_c00	4084840	SUCCEEDED	0	1	124	2019-09-24 05:25:09	2019-09-24 07:29:29	124
201302080000	post_low_p01	4084843	SUCCEEDED	0	1	141	2019-09-24 05:25:09	2019-09-24 07:45:20	140
201302080000	post_low_p02	4084845	SUCCEEDED	0	1	141	2019-09-24 05:25:09	2019-09-24 07:45:23	140
201302080000	post_low_p03	4084844	SUCCEEDED	0	1	140	2019-09-24 05:25:09	2019-09-24 07:45:04	140
201302080000	post_low_p04	4084846	SUCCEEDED	0	1	141	2019-09-24 05:25:09	2019-09-24 07:45:19	140
201302080000	post_low_c00	4084925	SUCCEEDED	0	1	120	2019-09-24 05:30:08	2019-09-24 07:30:10	120
201302080000	postacc	4085006	SUCCEEDED	0	1	1	2019-09-24 05:35:06	2019-09-24 05:36:23	1
201302080000	prdgen_low_p01	4085152	SUCCEEDED	0	1	125	2019-09-24 05:40:08	2019-09-24 07:44:59	125
201302080000	prdgen_low_p02	4085153	SUCCEEDED	0	1	125	2019-09-24 05:40:08	2019-09-24 07:45:00	125
201302080000	prdgen_low_p03	4085154	SUCCEEDED	0	1	125	2019-09-24 05:40:08	2019-09-24 07:44:43	125
201302080000	prdgen_low_p04	4085155	SUCCEEDED	0	1	125	2019-09-24 05:40:08	2019-09-24 07:44:55	125
201302080000	prdgen_low_c00	4085156	SUCCEEDED	0	1	110	2019-09		
201302080000	extractvars	4087466	SUCCEEDED	0	1	17	2019-09		
201302080000	post_track	4087467	SUCCEEDED	0	1	0	2019-09		
201302080000	enspost	4087469	SUCCEEDED	0	1	1	2019-09		
201302080000	keep_data	4087895	SUCCEEDED	0	1	0	2019-09		
201302080000	archive	4087897	SUCCEEDED	0	1	27	2019-09		
201302080000	cleanup	4088503	SUCCEEDED	0	1	1	2019-09		



Checking files in KEEP_DIR

(-5 means this folder does not exist!)

Folder Name	Number of Files in KEEP_DIR
f2d	520/520
f3d	520/520
ensstat	160/160
tctrack	-5/0-NO STORMS
logs	46

DATA Archive

There are 6 files in HPSS

gefs.20130208_00.ensstat.tar	Sep 24 08:05	204.6 MB
gefs.20130208_00.ensstat.tar.idx	Sep 24 08:05	83.2 KB
gefs.20130208_00.pgrb2p25.tar	Sep 24 08:30	139.9 GB
gefs.20130208_00.pgrb2p25.tar.idx	Sep 24 08:30	410.9 KB
gefs.20130208_00.pgrb2p5.tar	Sep 24 08:32	13.4 GB
gefs.20130208_00.pgrb2p5.tar.idx	Sep 24 08:32	124.2 KB

Real-Time Monitoring System for Retrospective

The Real-Time Monitor System for Retrospective

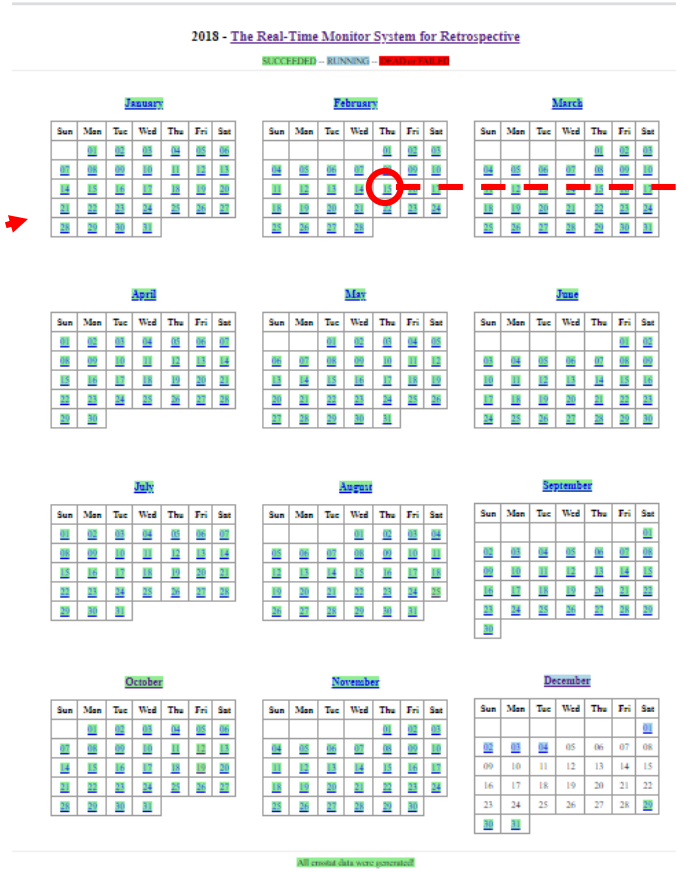
SUCCEEDED -- RUNNING -- DEAD or FAILED

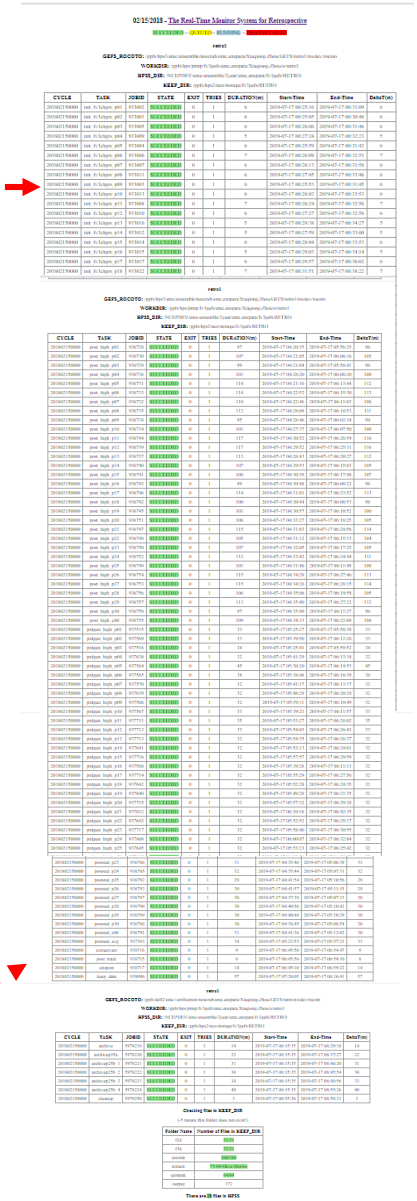
00UTC

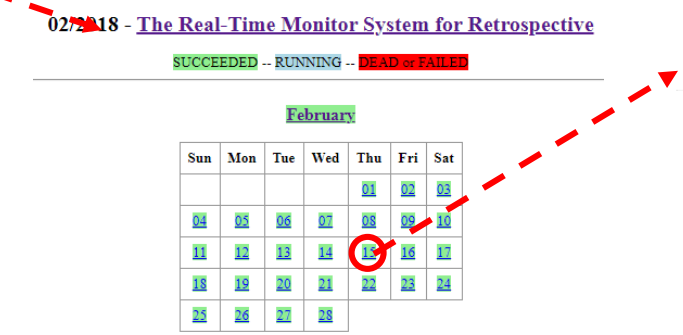
Year	Month										
2017					06	07	08	09	10	11	12
2018	01	02	03	04	05	06	07	08	09	10	11	12
2019	01	02	03	04	05	06	07	08	09	10	11	

SUMMER 12UTC

Year	Month										
2017						07	08	09	10		
2018						07	08	09	10		
2019						07	08	09	10		







Thanks!!!

Poster # 654: *Toward an Optimal Configuration of Dynamics and Physics for GEFS v12*

Post at 04:00 PM - 06:00 PM on Tuesday, January 14, 2020

J55.3: *Computational Resources Optimization in the NCEP Coupled Atmospheric Wave–Chemistry Global Ensemble Forecast System*

Present at 3:30 PM on Wednesday, 15 January 2020 Room 212