ESCAPE 2

Full-system sized ensemble forecasts within the URANIE framework

Dissemination Level: Public

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 800987.

www.hpc-escape2.eu

Funded by the European Union

Co-ordinated by ECMWF

Energy-efficient Scalable Algorithms for Weather and Climate Prediction at Exascale

Author **Michiel Van Ginderachter**

Date **30/09/2021**

Research and Innovation Action

H2020-FETHPC-2017

Project Coordinator: Dr. Peter Bauer (ECMWF)

Project Start Date: 01/10/2018

Project Duration: 36 month

Published by the ESCAPE-2 Consortium

Version: 1.0

Contractual Delivery Date: 30/09/2021

Work Package/ Task: WP4 / Task 4.6

Document Owner: RMI

Contributors: CEA, ECMWF

Status: Final

Table of Contents

Figures

Tables

Introduction

1.1 Background

ESCAPE-2 has developed world-class, extreme-scale computing capabilities for European operational numerical weather and climate prediction systems. It continued the pioneering work of the ESCAPE project. The project has attacked all three sources of enhanced computational performance at once, namely (i) developing and testing bespoke numerical methods that optimally trade off accuracy, resilience and performance, (ii) developing generic programming approaches that ensure code portability and performance portability, (iii) testing performance on HPC platforms offering different processor technologies.

ESCAPE-2 has prepared weather and climate domain benchmarks that allow a much more realistic assessment of application specific performance on large HPC systems than current generic benchmarks such as HPL 1 and HPCG 2. These benchmarks are specifically geared towards the pre-exascale and exascale HPC infrastructures that the European Commission and Member States will invest in through the European Commission Joint Undertaking.

ESCAPE-2 also combined generic uncertainty quantification tools for high-performance computing originating from the energy sector with ensemble-based weather and climate models to quantify the effect of model and data related uncertainties on forecasting –a capability, which weather and climate prediction has pioneered since the 1960s. This collaboration combined user-friendly tools from one community with scientific expert knowledge from another community and achieved economy of scales beyond the scope of each domain.

1.2 Scope of this deliverable

1.2.1 Objectives of this deliverable

Deliverable D4.6 aims to capitalize on the developments and insights of the previous tasks by implementing a typical operational ensemble system (HarmonEPS) in the URANIE framework. The goals of the deliverable are (i) integrate HarmonEPS in the URANIE framework; (ii) perform a typical VVUQ assessment for an entire ensemble prediction workload, with a focus on the sensitivity of the perturbation parameters; (iii) assess the computing performance of the full ensemble prediction system (EPS) chain.

1.2.2 Work performed in this deliverable

In this deliverable, the HARMONIE scripting system (based on EcFlow), used for running the ensemble predictions was adapted to integrate the tasks needed by the URANIE framework. For this, the typical URANIE ‘black-box approach’ – to run the model as black-box by only changing the input parameters – was abandoned and in contrast, the URANIE routines were embedded into the scripting system of the model framework from HARMONIE. A sensitivity study of 2m relative humidity on the surface field perturbations was performed to check whether the resulting model configuration was working. Next, the sensitivity of the spread of some basic atmospheric parameters (2m temperature, 2m relative humidity, 10m wind speed, 500hPa geopotential height) on the correlation length scale of different perturbations that are added to the model simulations to generate a physically consistent ensemble spread was studied and the most influential correlation length scale was used in an optimization exercise. Finally, the different findings are discussed and some recommendations for future application and further development are made.

1.2.3 Deviations and counter measures

The content of task 4.5 of the original ESCAPE-2 proposal was changed, as it became apparent during the course of WP4 that (perturbations of) the initial state of NWP and climate models were not well suited for a potential VVUQ analysis through the URANIE platform. Instead, task 4.5, driven by a strong interest in calibration techniques by the NWP community, focused on implementing state-of-the art calibration techniques in the URANIE framework.

Consequently, this deliverable does not include a sensitivity study on the perturbations of the initial state but instead performs a calibration/optimization study of a typical perturbation input parameter.

2 Running the HARMONIE ensemble prediction system in the URANIE VVUQ framework

2.1 Introduction to the HARMONIE ensemble prediction system

The HARMONIE ensemble prediction system, also known as HarmonEPS is the limited-area, short-range, convection-permitting ensemble prediction system developed and maintained by the HIRLAM consortium (now merged with the ALADIN consortium to form the ACCORD consortium). The forecast model solves the non-hydrostatic Eulerian equations in a mass-based vertical-coordinate system with semi-implicit time stepping and semi-Lagrangian advection [1].

There exist two configurations for handling subgrid physical phenomena. The most-used one is the HARMONIE-AROME configuration, developed considering the deep convection as resolved [2]. The second configuration – HARMONIE-ALARO – has physics parametrization schemes that are optimised for use at multiple resolutions within the so-called “gray zone” where deep convection is partly resolved (2 - 10 km) [3]. In both configurations, the surface processes are modeled using SURFEX [4].

For weather predictions, it is not only important to forecast the most likely scenario for future weather. It is equally important to quantify the uncertainties of specific predictions. This is typically done via the use of ensemble methods for which a number of slightly different weather forecasts (ensemble members) are run in parallel to provide a distribution of possible weather scenarios. For ensemble predictions, it is important that the spread of the ensemble is a good estimate for the actual uncertainty of predictions and the tuning of ensemble spread is therefore important during the development of Ensemble Prediction Systems (EPSs). HarmonEPS can account for uncertainties in the initial conditions (both surface and upper air), lateral boundaries, and model description. While HarmonEPS has different options to represent the uncertainty in each category, we will limit ourselves here in presenting only the default options. For a more detailed explanation of all possible options, the reader is referred to [1].

Lateral boundary perturbations are naturally included when HarmonEPS is nested in a coarser-resolution or global EPS. The spread induced from the boundaries is then typically controlled by the selection of the coarser-model members.

The default perturbation strategy for upper-air initial conditions is to add perturbations from the nesting model using the corresponding boundary file at initial time to the HarmonEPS control member. When using the IFS ENS, the upper-air perturbations are simply the differences between the IFS ENS members and the IFS ENS control at initial time.

Surface perturbations are applied to parameters in the SURFEX analysis after the surface data assimilation is completed. Perturbations for each parameter and member are created by generating an independent field of white noise and applying a recursive Gaussian filter until the prescribed correlation length is reached. The random noise field is then clipped to the range ±2 and scaled depending on the parameter [5]. Perturbations can be multiplicative or additive depending on the parameter. Table 1 shows the standard deviation and type of scaling applied for each of the perturbed parameters.

Model uncertainty can be modelled by running HarmonEPS in a multi-physics setup. Each ensemble member in HarmonEPS then has a unique combination of physics parametrizations. There is even the option to build a multi-model ensemble from HARMONIE-AROME and HARMONIE-ALARO with differences in both physics and dynamics. Finally, two perturbation schemes are available to represent the model uncertainty. HarmonEPS has the possibility to run with perturbed parametrized tendencies based on the adapted version [6] of the SPPT scheme [7]. HarmonEPS has also the option to use the stochastically perturbed parametrizations (SPP) scheme based on [8], where the parametrization parameters are gradually changed during the forecast depending on space and time. Both perturbation schemes are discussed in more detail in section 4.

Parameter	Short name	Standard deviation	Perturbation type
Vegetation fraction	VEG	0.1	Multiplicative
Leaf area index	LAI	0.1	Multiplicative
Thermal coefficient of vegetation	CV	0.1	Multiplicative
Vegetation Surface roughness length over land	Z0	0.2	Multiplicative
Albedo	ALB	0.1	Multiplicative
Sea surface temperature	SST	0.25	Additive
Soil temperature	ST	1.5	Additive
Soil Moisture	WG	0.1	Multiplicative
Snow Depth	SNOW	0.5	Multiplicative
Surface fluxes of sea	NA	0.2	Multiplicative

Table 1: Perturbed SURFEX surface variables, with their respective standard deviation and perturbation type.

For a more in-depth description of HarmonEPS and the impact of the different perturbation techniques on the ensemble skill, the reader is referred to [1] and the references therein.

2.2 Running HarmonEPS using the URANIE platform

The URANIE platform is an open-source framework developed at the Alternative Energies and Atomic Energy Commission (CEA) in order to deal with uncertainty propagation, surrogate models, optimization issues and code calibration [9]. The URANIE platform was successfully used for a sensitivity study of the shallow-water equations (see [D4.2]) and sensitivity study and uncertainty quantification of the ACRENEB2 radiation dwarf (see [D4.4]). Here, making a rather large leap, the URANIE platform is used to perform a sensitivity analysis and calibration exercise of a typical operational ensemble forecasting system.

2.2.1 Integrating the URANIE framework and the HarmonEPS scripting system

Usually, URANIE considers the code or model under investigation as a black-box, requiring only a very limited amount of communication with the code of the model via input- and output-files. Given the complex nature of running an EPS, involving the staging and archiving of data and cycling of fields between different forecasts and assimilation cycles (see Fig. 1 for a schematic overview.), it was chosen not to consider HarmonEPS as a black-box, but on the contrary to integrate the individual URANIE tasks (e.g. selecting the input-parameter distributions, setting the namelists, gathering the output, etc.) in the HarmonEPS scripting system.

Figure 1: Schematic overview of all the different steps involved in a typical HarmonEPS ensemble forecast. All jobs and dependencies are managed by the EcFlow workflow package. Note that the staging and preparation of the boundary files (MakeCycleInput) is completely separated from the actual forecasts (Date), preventing a simple 'black-box' approach typically used in URANIE.

The new workflow can be roughly described as follows: First URANIE is initialized. This means that the desired experiment (sensitivity study, calibration…) is created in URANIE, the number of model evaluation iterations (here all forecasts of all ensemble members in the selected period) is calculated and the values of the parameters or variables under study are set for each specific iteration. In standard URANIE applications, the model would be run from within the URANIE program. Here, all information is stored in an URANIE-readable file for later use and URANIE is exited and restarted after the next model iteration.

In the next step, HarmonEPS stages all the data needed for the full forecasting period (MakeCycleInput in Fig. 1). Then one URANIE iteration of a full forecasting period is run, including the assimilation cycle and postprocessing of the output data, followed by the calculation and storage of the specific metrics or cost functions needed for the sensitivity study or calibration by URANIE (UranieMetric in Fig. 2).

This is repeated for each iteration given by URANIE. Finally, when all iterations are finished, URANIE reads in all the data needed and calculates the statistics, depending on the experiment created. Special care was given to the new iterative process in HarmonEPS such that experiments with an a priori unknown number of iterations such as calibration experiments can also be accommodated.

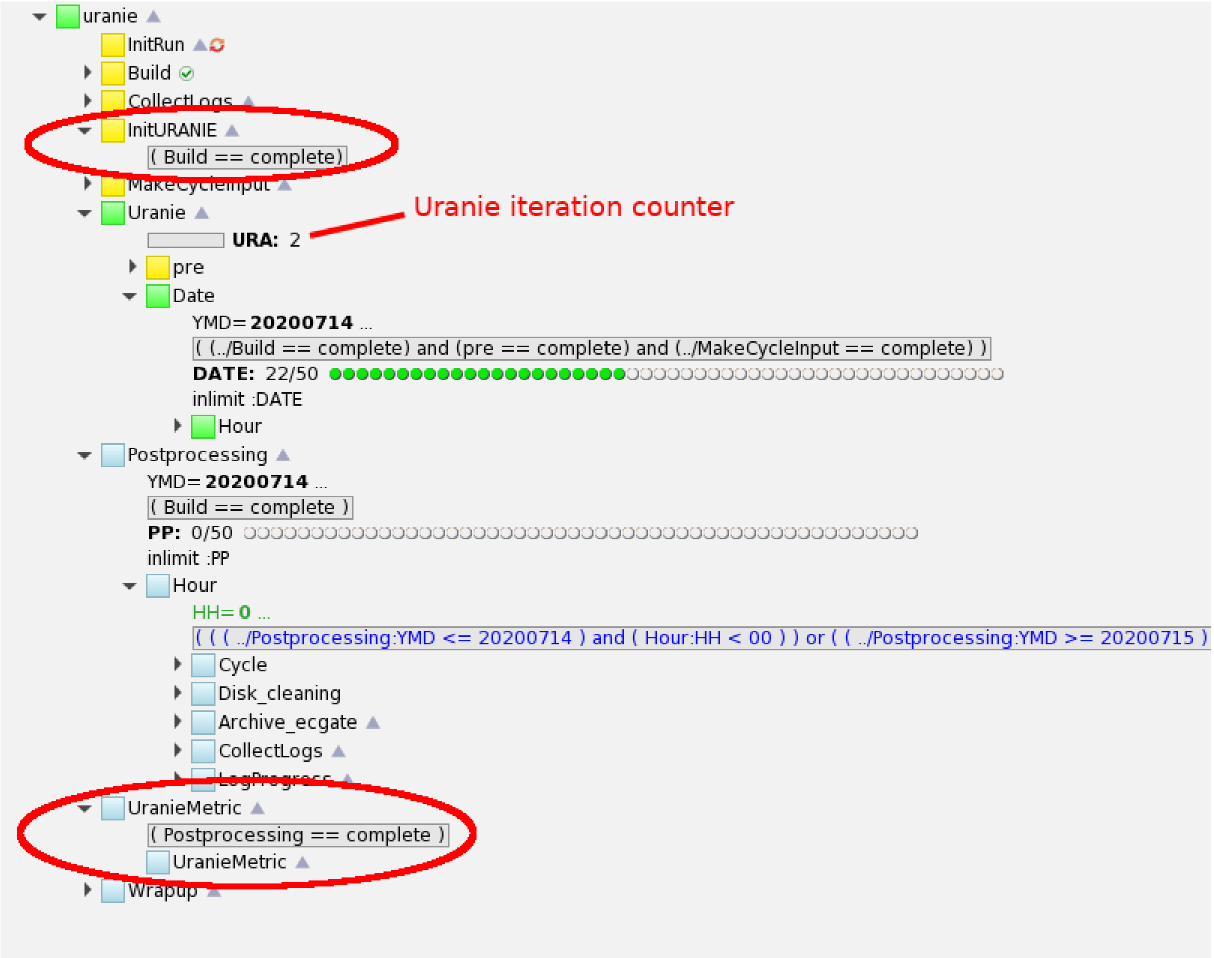

Figure 2: The new EPS workflow, with the new steps involving URANIE tasks marked in red.

2.2.2 General HarmonEPS configuration

For all experiments discussed in this deliverable HarmonEPS runs the default HARMONIE-AROME configuration with a grid-spacing of 2.5 km, 65 vertical levels and a time-step of 60 seconds. The domain is centered over Western Europe, ranging from 1°E to 17°W and 45.5°N to 57.5°N (see Fig. 3). The control member runs a surface and upper air data assimilation cycle of 3 hours. 10 ensemble members are created by starting from the control analysis, adding either only surface perturbations or tendency and parameter perturbations, depending on the experiment, thus creating an ensemble of 11 members. The boundary conditions are provided by IFS HRES. Prior to starting the sensitivity study or calibration with URANIE, the assimilation cycle was allowed to spin-up for 13 days, starting at 1 July 2020. The period under investigation covers only two days, where for each day a 36h forecast is performed at 0000 UTC.

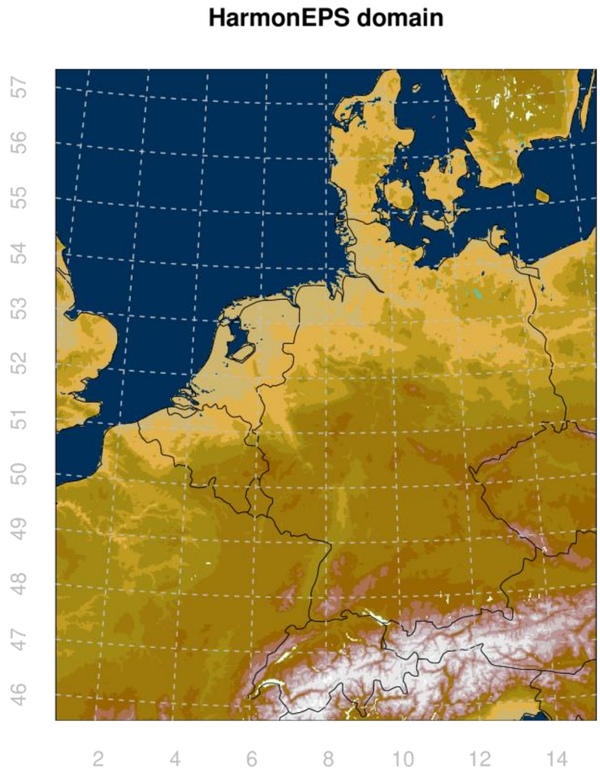

Figure 3: The HarmonEPS domain used for all experiments described here.

3 Sensitivity of RH2m BIAS on the perturbed surface variables

3.1 Background

While the results presented in [1] show that HarmonEPS performs quite well compared to IFS ENS, in-depth analysis of HarmonEPS forecasts reveals a problem with the 2m relative humidity (RH2m). The problem is clearly illustrated in Fig. 4, where all perturbed members are drier than the control member (black). Further analysis, running many different HarmonEPS configurations revealed the dry bias of the perturbed members to be region dependent and the source of the problem was identified to be related to the perturbation of the soil moisture (WG) [10].

This methodology of *guestimating* the source of the problem, reconfiguring and running the EPS by hand, investigating the results and finally validating the hypothesis is a tedious, complex and computationally demanding task.

Estimating the sensitivity of an output variable on different parameters is a perfect use-case for the Morris screening method (explained below) included in the URANIE package. Therefore, as a first sanity test of the new HarmonEPS-URANIE setup we will try to confirm the conclusion drawn in [10], identifying the soil moisture perturbations as the source of the dry bias seen in the perturbed ensemble members.

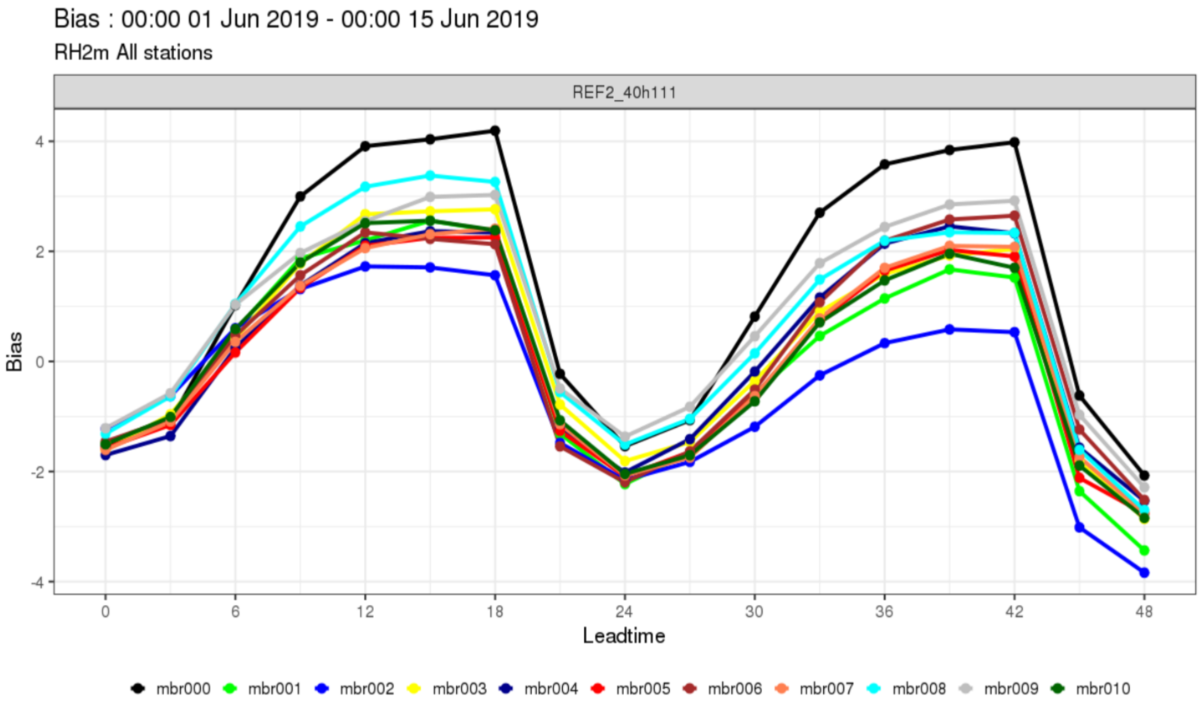

Figure 4: 2m relative humidity bias for every member of the HarmonEPS over a 15-day period (1 June - 15 June 2019). Taken from [10]. A perfect prediction would be unbiased and would therefore be a flat line at 0.

3.2 The Morris Method

In this experiment the Morris screening method is used to investigate the sensitivity of RH2m to the different perturbed surface parameters.

The Morris method is an effective screening procedure and a more robust version of the One-factor-At-a-Time (OAT) variation method. Instead of varying every input parameter only once, the Morris method repeats the OAT principle several (*r*) times. For each iteration (called a trajectory) a different randomly chosen starting point in the input parameter space is used. For every parameter variation in trajectory *t*, the elementary effect (EE) is computed:

$$EE_i^t = EE_i(\boldsymbol{X}^t) = \frac{y(X_1^t,\dots,X_i^t,\dots,X_n^t) - y(X_1^t,\dots,X_i^t + \Delta_i^t,\dots X_n^t)}{\Delta_i^t}.$$

y is the output metric under investigation, $X_1 \dots X_n$ are the input parameters under investigation, and Δ_i^t is the chosen variation for input parameter i in the trajectory t. The recommended value for the variation is $\Delta = \frac{p}{2(p-1)}$. *p* is called the level, the chosen number of intervals this range should be split into [11]. With the repetition of this procedure *r* times, resulting in $r(n_X + 1)$ code evaluations, it is possible to compute basic statistics on the elementary effects computed for every input parameter, as

$$\mu_i^* = \frac{1}{r}\sum_{t=1}^{r} |EE_i^t|,$$

$$\sigma_i^2 = \frac{1}{r-1}\sum_{t=1}^{r} (EE_i^t - \mu_i^*)^2.$$

Depending on the (μ_i^*, σ) values, the inputs can be sorted into different categories:

- Input has negligible effect on the output: both μ_i^* and σ are small.
- Input has a linear effect without interaction with other inputs: μ_i^* is large but σ is small.
- Input has a non-linear effect and/or interaction with other inputs: both μ_i^* and σ are large.

3.3 Experimental setup

Here, we have chosen the perturbation standard deviations of the first 9 variables (sea surface flux perturbations were turned off) in Table 1 as the inputs for the sensitivity study. Considering the computational cost of one iteration (two 36h ensemble forecasts of 11 members with 3h assimilation cycle), the number of trajectories r was kept reasonably small and set to 4, resulting in $r(n_X + 1) = 40$ full ensemble forecast evaluations. For each input, the range was set to $\pm$10% of its default value (third column in Table 1) and the level *p* was set to 10, the URANIE default value.

The ensemble forecasts use the default configuration explained in section 2.2.2 with the exception that only the surface perturbations are active. No other source of perturbations (IC, LBC, SPPT or SPP) is used.

After each ensemble forecast iteration the RH2m bias of each ensemble member with respect to the control member is calculated and averaged over the domain and over all ensemble members. This averaged bias is then used as output in the Morris screening method and used for calculating the elementary effects.

3.4 Results and discussion

The results of the Morris sensitivity experiment are summarized in Fig. 5. This figure shows the locations of all 9 inputs in the (μ^*, σ)-plane. The RH2m bias is only sensitive to the standard deviation of 4 perturbed variables, i.e. snow depth (SNOW), soil temperature (ST), sea surface temperature (SST) and soil moisture (WG). From these 4 inputs, the snow depth and soil temperature standard deviation have both small $\mu^*$and σ, indicating small interaction with other inputs as well as a small effect on the bias. Sea surface temperature standard deviation has similar σ but larger μ^*. Looking at the relatively large value of μ^* for soil moisture (WG), it seems the RH2m bias is most sensitive to the soil moisture standard deviation. This parameter also shows most interaction with the other inputs.

These results identify the soil moisture as the parameter with the dominating effect on the RH2m bias, confirming the analysis made in [10]. However, the results here were found using a systematic and fully automated approach, contrary to [10] where the different configurations were selected based on expert-knowledge and reconfigured *by hand.* Furthermore, no expert knowledge of the inner-workings of the surface perturbation scheme was needed for the experiment performed here.

Finally, we address the parameters (CV, VEG, ALB, LAI and Z0) that show no influence on the RH2m bias. The lack of sensitivity on albedo (ALB) and roughness length (Z0) is explained by the fact that these fields are only updated and thus perturbed every 10 days. Such an update is probably not performed in the 2-day forecast window used here.

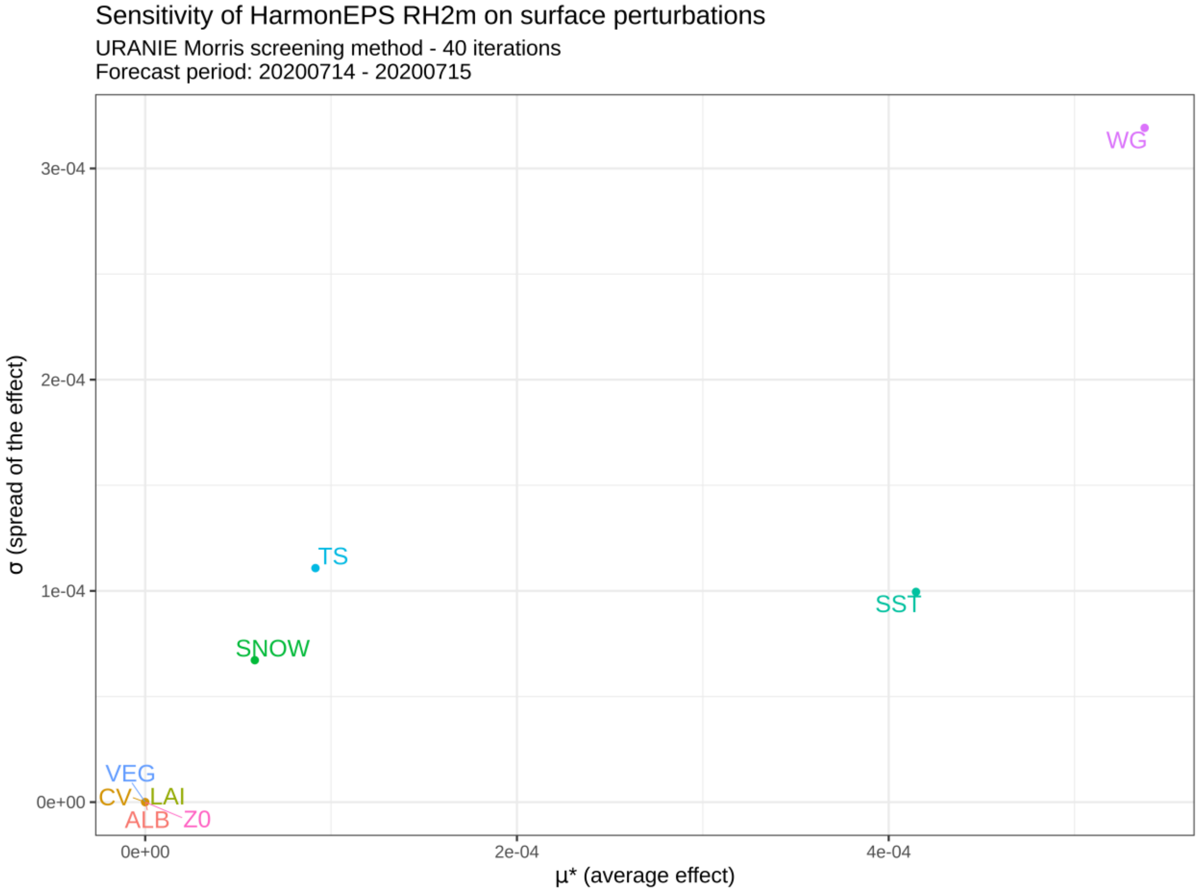

Figure 5: Morris screening indices of the vegetation fraction (VEG), leaf area index (LAI), thermal coefficient of vegetation (CV), vegetation surface roughness length over land (Z0), albedo (ALB), sea surface temperature (SST), soil temperature (ST), soil moisture (WG) and snow depth (SNOW) perturbation standard deviations for the RH2m BIAS.

For the remaining three parameters (CV, VEG and LAI) no immediate explanation for the zero-sensitivity was found by the authors. However, in the months after the analysis we were informed that independent experiments by the HarmonEPS development team had revealed a bug in the surface perturbation code, causing the physiography related parameters CV, VEG and LAI to not be perturbed even when the surface perturbation scheme is switched on.

So besides identifying the soil moisture as the probable source of the drying of the perturbed ensemble members, the systematic URANIE screening method was also able to, unbeknownst to the authors, expose a bug in the code.

Finally, we discuss here briefly the computational performance of the full HarmonEPS + URANIE workload. As expected, most computationally expensive are the iterations of the full forecast cycle. With one iteration taking between 8 and 12 hours using 324 CPUs, depending on the of the load the machine. Therefore, sensitivity experiments demanding more iterations such as a Sobol analysis [12] can become hard to realise, especially when the machine has to be shared with many users.

The computational time needed for URANIE preparation and URANIE Morris analyzing tasks, however, are negligible compared to the total CPU-time of the workload.

4 Sensitivity study and optimization of the SPPT and SPP correlations length scale

4.1 Background

As mentioned in Section 2.1 HarmonEPS includes two perturbation schemes for representing the model uncertainty, namely the SPPT and SPP scheme. The former perturbs all net physics tendencies with random 2D multiplicative noise. For each ensemble member, the perturbations are drawn from a different realization of a pattern generator with spatial and temporal correlations. In the latter 14 parameters (7 cloud, 4 radiation, and 3 turbulence parameters), identified by physics experts as most uncertain, are perturbed by drawing the multiplicative perturbations independently (mostly from a log-normal distribution) for each parameter and ensemble member using the same pattern generation as described above. Finally, perturbations are clipped in order to stay within the physical range defined by the experts.

In HarmonEPS, the pattern generator originally used for SPPT in [7] was replaced by the stochastic pattern generator (SPG) developed in [13] due to a mismatch between the specified and resulting pattern characteristics [1]. This pattern generator (but also the one originally used in SPPT) has three main parameters to be set: The standard deviation of the pattern generator σ, The horizontal length scale L and time scale of decorrelation τ. For HarmonEPS the default values used are: $\sigma = 0.3, L = 200\ \mathrm{km}$ and $\tau = 8\mathrm{h}$ for SPPT, while $\sigma = 3.0, L = 200\ \mathrm{km}$ and $\tau = 12\mathrm{h}$ are used for the SPP pattern generator.

These values are typically either the result of manual time-consuming tuning exercises [e.g. 13] or arbitrarily chosen [e.g. 6]. Here such a manual tuning exercise is replaced by an automated optimization exercise using the Efficient Global Optimization (EGO) routine of the URANIE platform. A more detailed explanation about the used optimization technique is given below.

Here, we first of all want to study the feasibility of using URANIE with an EPS for optimization rather than finding the optimal value for all pattern generator parameters. Therefore, only a simple setup is used where only the correlation length scale for either SPP or SPPT is optimized with respect to a cost function using only one verification metric for one typical meteorological variable. The choice for the length scale is arbitrary, an identical experiment for the time scale or standard variation would be as valuable.

In order to determine which correlation length scale has the highest impact on which variable, a sensitivity study similar to the one described above is performed in section 4.2. The most sensitive correlation length scale for all combinations of variables is then used in the optimization study.

4.2 Sensitivity study

4.2.1 Experimental setup

The Morris screening method is again used to perform the sensitivity study. Here, the inputs are the correlation length scales L_{SPPT} and L_{SPP} of the SPPT and SPP pattern generator respectively. As we are most interested on the effect the changes in length scales have on the ensemble as a whole rather than on the individual members, the outputs are defined as the ensemble spread of a selection of four variables, i.e. the ensemble spread of 2m temperature (σ_{T2m}), 2m relative humidity (σ_{RH2m}), 10m wind speed (σ_{S10m}) and 500hPa geopotential height(σ_{G500}). Given the small number of inputs used here (i.e. 2), the number of trajectories can be increased to $r = 10$, resulting in 30 full ensemble forecast evaluations. The level was kept at $p = 10$.

For both length scales the boundaries were first set at [50km, 300 km]. However, this caused crashes in the pattern generator. By Increasing the lower bound to 100km we were able to prevent these crashes.

The ensemble forecast again uses the default setup, explained in section 2.2.2 with the adaption that only SPPT and SPP perturbation schemes are active, meaning there are no IC, LBC or surface perturbations.

4.2.2 Results and discussion

Figure 6 shows the results of the Morris screening. For all variables the spread is most sensitive to changes in the SPP correlation length scale L_{SPP}. These results are in line with the work presented in [14], where it is shown that the spread added by SPP is typically larger than the spread added by SPPT when using the default parameters for both schemes.

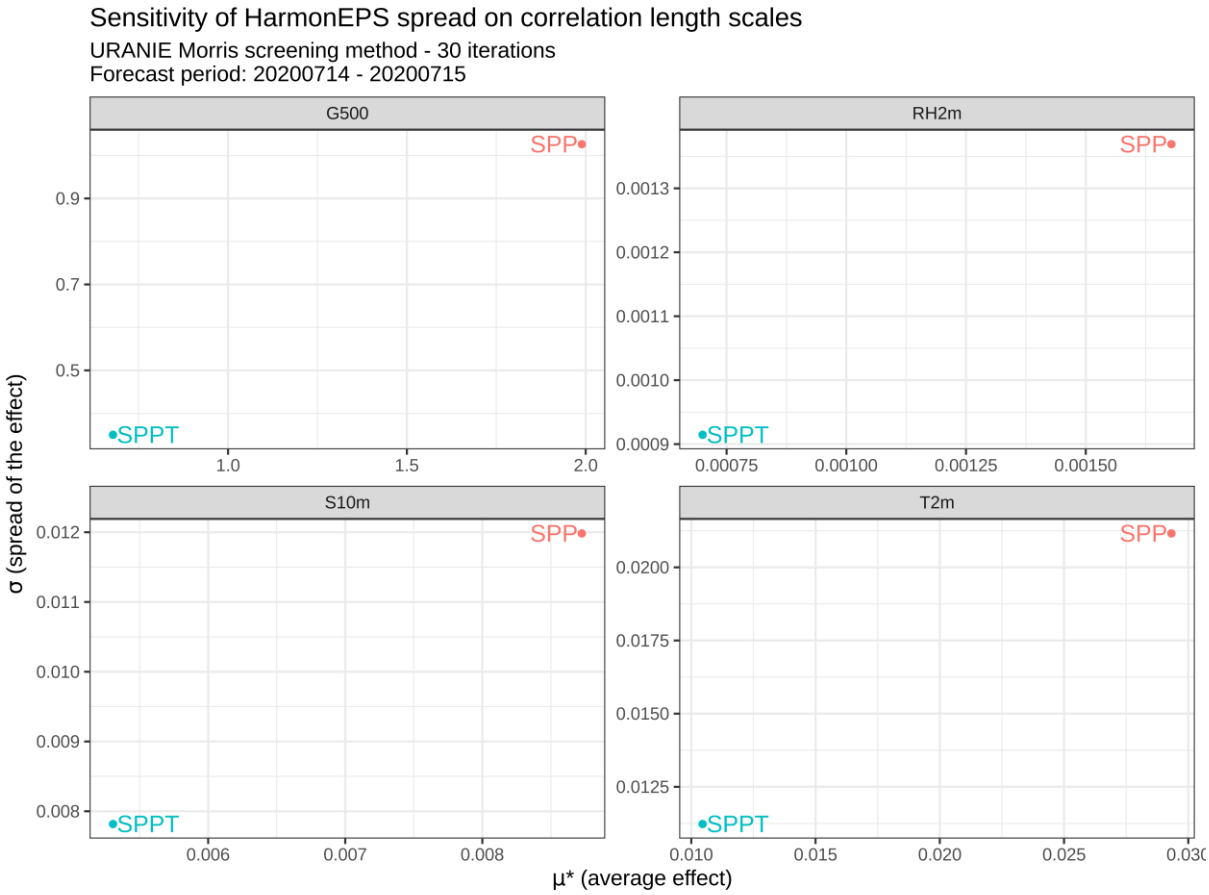

Figure 6 Morris screening indices of L_{SPPT} and L_{SPP} for the ensemble spread of 500hPa geopotential (G500), 2m relative humidity (RH2m), 10m wind speed (S10m) and 2m temperature (T2m).

Changes in L_{SPP} have not only the largest impact on the spread, the larger σ indices indicate that the impact is also more non-linear. This is explained by the fact that SPPT perturbations are added at the end of the physics calculations to the net physical tendencies, preventing feedbacks between the different parameterizations. The SPP perturbations on the other hand are added to the parameters inside the parametrizations. As a consequence, different parameterizations might impact each other and perturbations might interact within a timestep.

To determine which variable's spread is most sensitive to changes in L_{SPP} we have defined the relative mean elementary effect as:

$$100\times\frac{\mu^*}{< SPREAD >_{\text{all iterations}}}$$

This relative mean elementary effect is shown in Table 2 together with the ensemble spread averaged over the 30 URANIE iterations for each variable.

Variable	Averaged Ensemble Spread	Relative effect of L_{SPP} $(100\times\mu^*_{SPP}/spread)$
T2m	0.94 K	3.13%
RH2m	0.056	3.00%
S10m	0.76 m s^{-1}	1.14%
G500	10.18 m	19.54%

Table 2 For each variable the ensemble spread, averaged over the 30 URANIE iterations of the ensemble forecast used for the Morris screening and the relative mean elementary effect.

From Table 2 it is clear that the changes in the correlation length scale of the SPP pattern generator have the largest impact on G500. Ideally, we would like to perform the optimization of L_{SPP} using a metric based on G500. Unfortunately, observations of this variable are not readily available in the HarmonEPS environment. Therefore, to avoid the additional work of gathering and postprocessing the 500 hPa observations, we will use the T2m in the following optimization exercise. Observations of T2m are readily available as they are extracted in the postprocessing step of HarmonEPS (see Fig. 1).

4.2.3 Calibration/Optimization technique

In the context of this study, the optimization must deal with the fact that the evaluation of the model is performed outside the URANIE algorithm (see Section 2.2.1). When the time-consuming ensemble forecast is performed an additional post treatment to extract the output values of interest is needed.

The Efficient Global Optimization, detailed in [15], also called Bayesian optimization meets the above criteria and only requires a reasonable amount of code evaluations, a great benefit when evaluating a time-consuming EPS. This method is based on the construction of a kriging surrogate (Gaussian process) and an adaptive strategy that assumes a compromise between the global improvement of the quality of kriging and the finding of an optimum. This approach identifies a new data set that is supposed to improve the optimization. The code evaluation and post treatment can be performed outside the algorithm after which the algorithm continues the optimum search.

The EGO is organized as follows:

1. An initial sample of the parameter under optimization is generated and the code is evaluated for each configuration.

2. A kriging surrogate (gaussian process) is trained on the results of the sampling.
3. Based on the kriging variance, the Expected Improvement Criteria (Definition is given in [15]) is optimized to identify the next configuration to evaluate for improving the optimization.
4. The code is evaluated and the results are added to the sample.
5. The stopping criteria, that indicate when the optimization must be stopped, are checked. If they are not satisfied, the algorithm is repeated starting at step 2.

Note that for EGO, the stopping criteria are mainly the number of evaluations that you can performed considering the time and CPUs that a single iteration requires.

4.2.4 Experimental setup

For the optimization exercise performed here, the HarmonEPS configuration is kept identical to the sensitivity experiment described above (Section 4.2.1). In the following, the SPPT pattern generator parameters are fixed at $\sigma_{SPPT} = 0.3, L_{SPPT} = 200km$ and $\tau_{SPPT} = 8h$. However, the correlation length scale L_{SPP} of the SPP pattern generator will be optimized and the remaining two SPP parameters were kept at their default values ($\sigma_{SPP} = 3.0$, $\tau_{SPP} = 12h$).

The cost function used in the optimization is the continuous ranked probability score (CRPS) [16] defined as the quadratic measure of discrepancy between the forecast CDF noted F and $\mathbb{I}(\mathrm{x} \geq \mathrm{y})$, the empirical CDF of the scalar observation y:

$$CRPS = \int_{-\infty}^{+\infty} [F(x) - \mathbb{I}(x \geq y)]^2 \mathrm{d}x,$$

with $\mathbb{I}$ the indicator function. The CRPS combines in a natural way a cost for the ensemble spread (-error) and a cost for the error in ensemble mean and is therefore an ideal candidate for the correlation length scale optimization cost function.

The CRPS of different variables can be combined into a single cost function and the used optimization even allows for multiple cost functions. Here, however, we have chosen, as explained in the previous section, to keep things simple and define the cost function as the CRPS of the 2m temperature only.

The optimization algorithm is initialized by randomly sampling 10 L_{SPP}values using the URANIE Latin hypercube sampling (LHS) [17] algorithm. For each value of L_{SPP} the full 2-day ensemble forecast cycle is run and the T2m CRPS is calculated. This input is used by the optimizer to propose a new correlation length scale L_{SPP}, for which a new forecast cycle is run. Ideally, this process is iterated until convergence of the correlation length scale is reached.

Observations for calculating the CRPS are limited to surface synoptic observation (SYNOP) stations and are extracted from the ECMWF mars database by the HarmonEPS postprocessing routines. Also, the interpolation of the forecast data to the SYNOP station locations is automatically performed by the HarmonEPS postprocessing.

4.2.5 Results and discussion

Figure 7 shows the evolution of both L_{SPP} and the CPRS of T2m during the successive iterations of the EGO. The first 10 iterations represent the LHS used to initialize the EGO. In iteration steps 11 and 12, the EGO explores the lower and upper limit respectively to improve the kriging surrogate. After 22 iteration steps an optimum $L_{SPP} \approx 230\ km$ is found. This is reasonably close to recommended value of 200 km, which was found by the HarmonEPS development team.

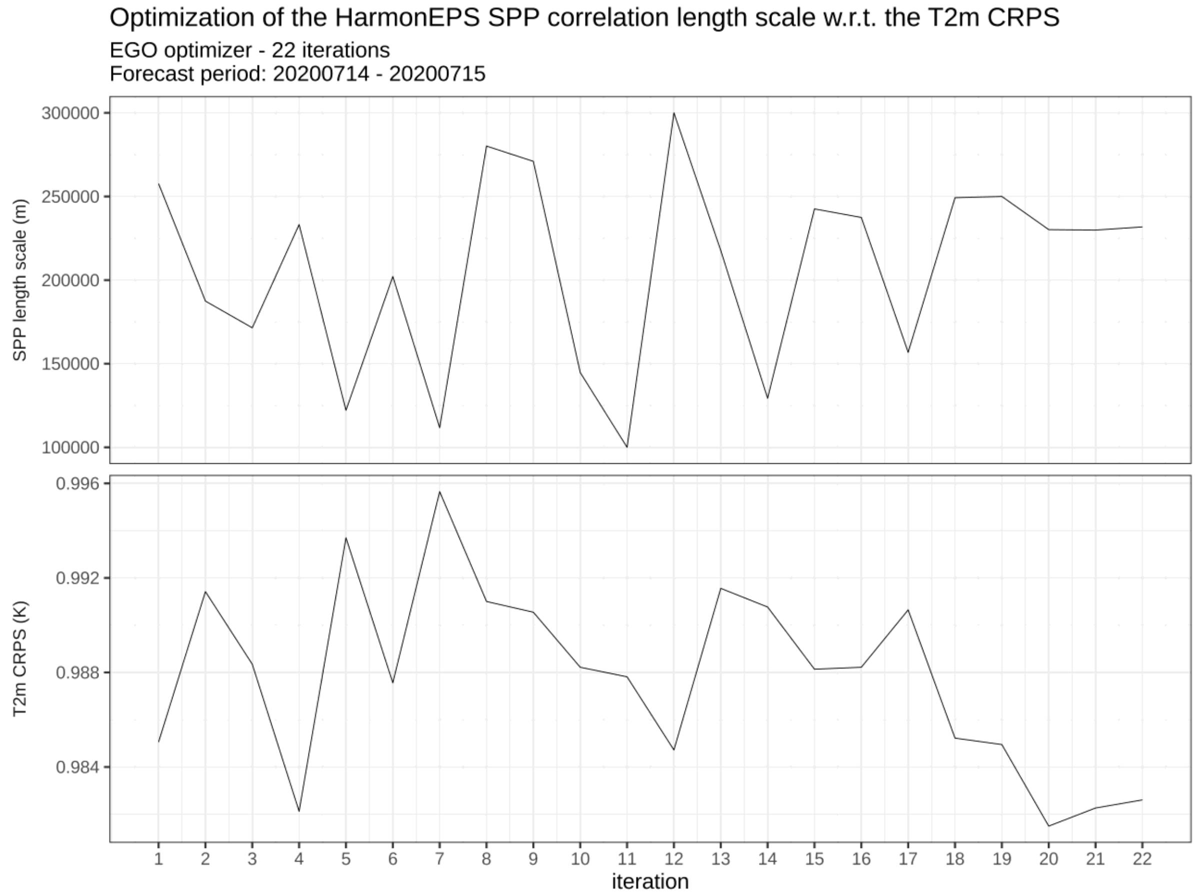

Figure 7: Evolution of L_{SPP} (top) and T2m CRPS (bottom) during the EGO optimization. The first 10 iterations represent the Latin Hypercube sampling. In the last 3 iterations we can see the convergence towards the optimal value.

Obviously, the results of this (proof-of-concept) tuning exercise should not be extended to a general context. Also, the changes in CRPS during the optimization are rather small, indicating that the direct influence of L_{SPP} on the CRPS is limited. For more general results, the cost function should include a broader selection of variables. In addition, all parameters of the both pattern generators (SPP and SPPT) should be tuned simultaneously in order to capture all relationships between the perturbation parameters. This is also discussed further in the next section.

5 Conclusion and outlook

In this work the feasibility of combining the URANIE VVUQ platform with the state-of-the art ensemble prediction system HarmonEPS was investigated. Combining both scripting systems was done by abandoning the typical URANIE black-box approach and instead integrating the typical URANIE routines inside the HarmonEPS scripting system. Although some intermediate knowledge of internal organization of HarmonEPS is needed[1], the URANIE integration was reasonably straightforward.

First a sensitivity study of the T2m BIAS on the perturbation standard deviations of different surface variables was performed. This study identified the soil moisture perturbation standard deviation as most influential, confirming results from earlier studies and ensuring that the URANIE integration in HarmonEPS was done correctly.

Next a calibration/optimization was performed. First the SPP pattern generator correlation length scale and T2m were identified as the optimal input – output combination through a sensitivity study. Finally, this input-output was used in a-proof of-concept optimization exercise using the URANIE EGO optimization algorithm. The outcome of the optimization exercise yields a sensible outcome in line with default HarmonEPS SPP settings.

The work done here focused mainly on the uncertainty quantification part of a typical VVUQ assessment since in the NWP and Climate community verification is best done by comparison with observations and for validation better-suited techniques for climate and NWP codes are available [e.g. 18 and D3.2] than those present in URANIE. Additionally, we would also like to mention that here all experiments were kept as simple as possible and forecast periods as short as possible given the a-priori unknown length and outcome of all experiments and the limited duration of the ESCAPE 2 project.

The results of this deliverable show that the automated tuning of model parameters can be performed successfully via a combination of the URANIE tool from CEA and a weather forecast model for full-grown ensemble prediction systems. The numerical tests are very demanding in terms of computational power as there are a number of dimensions for which it is known that an increase in the numbers would lead to better and more robust results: number of ensemble members N, number of start dates M, number of optimization iterations L, and the number of days of the forecasts O, with the cost of the optimization procedure scaling roughly with NxMxLxO. For this deliverable, the numerical tests were kept comparably simple with a single parameter that was optimized and relatively low numbers for N, M, L and O. However, the results clearly show that more can be done if more computational resources are spent. An optimization procedure that would, for example, tune the three length-scales and the three amplitudes of the SPPT scheme as used in the global IFS model will most likely require at least 10 forecast start dates, at least 10 ensemble members, more than 50 iterations, and 10 forecast days already reaching 50,000 forecast days that need to be simulated. If a high spatial resolution is targeted – such as the operational resolution of the ensemble forecast at ECMWF – such an optimization could quickly fill a large fraction of a EuroHPC supercomputer (based on previous work [19]). If a more optimal configuration of model parameters for ensemble predictions is found during this exercise, this could have a direct impact on the quality and uncertainty quantification of global weather prediction. In principle, all tools that are developed within this deliverable would be scalable to this level of complexity.

Considering the simple setup of the experiments performed here, the logical next step is to extend the forecasting period in order to increase the robustness of the results.

[1] The authors would like to express many thanks to Inger-Lise Frogner, who helped us with the initial setup of HarmonEPS and helped us with any issues that arose during the project.

The optimization technique used here allows for multiple inputs and outputs. Therefore, it would be interesting to extend the optimization to all SPPT and SPP parameters and optimize for a selection of prognostic variables. Finally, with more computing power it would be interesting to test the more iteration-demanding uncertainty quantification algorithms of URANIE.

6 References

[1] Frogner, I.-L., Andrae, U., Bojarova, J., Callado, A., Escribà, P., Feddersen, H., Hally, A., Kauhanen, J., Randriamampianina, R., Singleton, A., Smet, G., van der Veen, S., Vignes, O., 2019: HarmonEPS -The HARMONIE Ensemble Prediction System. , Weather and Forecasting, Vol. 34, Issue 6, 1909-1937, American Meteorological Society, https://doi.org/10.1175/ WAF-D-19-0030.1.

[2] Bengtsson, L., and Coauthors, 2017: The HARMONIE–AROME model configuration in the ALADIN–HIRLAM NWP system. Mon. Wea. Rev., 145, 1919–1935, https://doi.org/10.1175/MWR-D-16-0417.1.

[3] Termonia, P., and Coauthors, 2018: The ALADIN system and its canonical model configurations AROME CY41T1 and ALARO CY40T1. Geosci. Model Dev., 11, 257–281, https://doi.org/10.5194/gmd-11-257-2018

[4] Masson, V., and Coauthors, 2013: The SURFEXv7.2 land and ocean surface platform for coupled or offline simulation of earth surface variables and fluxes. Geosci. Model Dev., 6, 929–960, https://doi.org/10.5194/gmd-6-929-2013.

[5] Bouttier, F., Raynaud, L., Nuissier, O. and Ménétrier, B., 2016: Sensitivity of the AROME ensemble to initial and surface perturbations during HyMeX. Q.J.R. Meteorol. Soc., 142: 390-403. https://doi.org/10.1002/qj.2622

[6] Bouttier, F., Vié, B., Nuissier, O., & Raynaud, L., 2012: Impact of Stochastic Physics in a Convection-Permitting Ensemble, Monthly Weather Review, 140(11), 3706-3721.

[7] Buizza, R., Milleer, M. and Palmer, T.N., 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. Q.J.R. Meteorol. Soc., 125: 2887-2908. https://doi.org/10.1002/qj.49712556006

[8] Ollinaho, P., Lock, S.-J., Leutbecher, M., Bechtold, P., Beljaars, A., Bozzo, A., Forbes, R.M., Haiden, T., Hogan, R.J. and Sandu, I., 2017: Towards process-level representation of model uncertainties: stochastically perturbed parametrizations in the ECMWF ensemble. Q.J.R. Meteorol. Soc., 143: 408-422. https://doi.org/10.1002/qj.2931

[9] Blanchard, J.-B., Damblin, G., Martinez, J.-M., Arnaud, G., Gaudier, F., 2019: The Uranie platform: an open-source software for optimisation, meta-modelling and uncertainty analysis. EPJ Nuclear Sci. Technol., 5(4). https://doi.org/10.1051/epjn/2018050

[10] Harold, Mc Innes (2020). Drying ensemble members in HarmonEPS, Virtual EPS Presentations day, Monday 23 November 2020. https://hirlam.org/trac/wiki/Meetings/EPS/EPS2020-2

[11] Saltelli, A., Tarantola, S., Campolongo, F. and Ratto, M., 2002: The Screening Exercise. In Sensitivity Analysis in Practice (eds A. Saltelli, S. Tarantola, F. Campolongo and M. Ratto). https://doi.org/10.1002/0470870958.ch4

[12] Sobol, I.-M., 1993: Sensitivity indices for nonlinear mathematical models. Mathematical Modelling and Computational Experiment 1, 1993.

[13] Tsyrulnikov, M., and D. Gayfulin, 2017: A limited-area spatio-temporal stochastic pattern generator for simulation of uncertainties in ensemble applications. Meteor. Z., 26, 549–566, https://doi.org/10.1127/metz/2017/0815.

[14] Frogner, I.-L., Hally, A., Andrae, U., Hämäläinen, K., Kauhanen, J., Ollinahe P. (2020), Who's the winner? A comparison of different perturbations using tendency diagnostics. 42nd EWGLAM and 27th SRNWP Meeting, Online event. Wednesday 30th September, http://srnwp.met.hu/Annual_Meetings/2020/download/wednesday/PM/Frogner_EWGLAM2020.pdf.

[15] Jones D. R., M. Schonlau, and Welch W. J., 1998: Efficient Global Optimization of Expensive Black-Box Functions. Journal of Global Optimization, 13 (4), 455 – 492.

[16] Matheson J.E., Winkler R.L., 1976: Scoring rules for continuous probability distributions. Manag. Sci. 22(10):1087–1096, https://doi.org/10.1287/mnsc.22.10.1087

[17] Mckay, M. D., Beckman R. J., Conover W. J., 2000: A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code, Technometrics, 42:1, 55-61, https://doi.org/10.1080/00401706.2000.10485979

[18] Milroy, D. J., Baker, A. H., Hammerling, D. M., and Jessup, E. R.2018: Nine time steps: ultra-fast statistical consistency testing of the Community Earth System Model (pyCECT v3.0), Geosci. Model Dev., 11, 697–711, https://doi.org/10.5194/gmd-11-697-2018

[19] Tuppi, L. and Ollinaho, P. and Ekblom, M. and Shemyakin, V. and Jarvinen, H., 2020: Necessary conditions for algorithmic tuning of weather prediction models using OpenIFS as an example, Geosci. Model Dev., 13, 5799--5812, https://gmd.copernicus.org/articles/13/5799/2020/, 10.5194/gmd-13-5799-2020

[D3.2] Van Ginderachter M, 2019: D3.2: Assessment of the evaluation and verification tools based on the v0-benchmark configurations, ESCAPE-2 consortium, Work Package 3, Task 3.2, https://www.hpc-escape2.eu/resources/d32-assessment-evaluation-and-verification-tools-based-v0-benchmark-configurations

[D4.2] Chocat R., 2019: D4.2: Assessment of URANIE VVUQ performance for the Shallow water toy model, ESCAPE-2 consortium, Work Package 4, Task 4.2.

[D4.4] Chocat R., Van Ginderachter M., Dueben P., Nielsen K.P., Paronuzzi S., Serradell K., 2020: D4.4: Assessment of URANIE VVUQ performance for the ACRANEB2 dwarf, ESCAPE-2 consortium, Work Package 4, Task 4.4

Document History

Version	Author(s)	Date	Changes
0.1	RMI	10/08/2021	First draft.
0.2	RMI/CEA/ECMWF	15/09/2021	Added section by CEA, Proofreading.
1.0	RMI/CEA/ECMWF	30/09/2021	Added reviewer remarks, added final results.

Internal Review History

Internal Reviewers	Date	Comments
DMI	28/09/2021	See email Kristian Pagh Nielsen.
LU	28/09/2021	See annotated document Mike Gillard.

Effort Contributions per Partner

Partner	Efforts
RMI	8
ECMWF	0.1
CEA	1
Total	**9.1**

ESCAPE 2

ECMWF Shinfield Park Reading RG2 9AX UK
Contact: peter.bauer@ecmwf.int