This is the second review of “SKRIPS v1.0: A regional coupled ocean–atmosphere modeling framework (MITgcm–WRF) using ESMF/NUOPC, description and preliminary results for the Red Sea”, by Sun et al.

General Comments

The validation of a number of the fields is not convincing. What is an anticipated diurnal oscillation from a 30-km reanalysis vs a 9-km model output supposed to look like? What impact is the coarsened resolution of the ERA5 near the coast, where the surface air temperature differences could be larger than 20 C with a grid cell (desert vs ocean)? The only single figure with any sort of statistical inference possible shows that the coupled model performs no better than the test specifically expected to perform poorly.

In general, a comparison against a physically unrealistic month-long constant SST is problematic. Other than possibly a single mention of CPL vs ATM.STA at the beginning of the paper, most of the comparisons should be between CPL and ATM.DYN.

There are too many instances of the authors using “may”, “perhaps”, “hypothesize”, etc. With a numerical model, all of these uncertain statements can be directly attributed.

There is no simply stated working definition of a heatwave. From the figures, a heatwave is not entirely obvious, so “capturing” a heatwave is quite subjective.

In several places the authors conclude a paragraph with a wrap up sentence to the effect that the CPL test performs well. In most of those examples of validations or comparisons, the other tests performed well also.

A couple of the scaling comments are incomplete, such as only talking about the total number of processes or the mention of overlap cells.

Several of the references to the figures refer to labels that do not exist. Some of the figure captions would be improved if they were more stand alone.

The paper would benefit from a good proofreading. There are misspellings, missing words, undefined terms, and a few unusual phrasings.

The authors do not make a strong case for their selection of this particular domain and the simulated event. This paper has as a focus a series of heatwave events where 84% of the domain is land (and desert). For the coastal temperature comparisons, there is no mention of possible sea breeze effects. This is not an ideal set up that benefits from coupled interactions between ocean and atmosphere.

Specific Comments

Page 4 line 2:

is shown in Fig. 1(a)

There is no (a) label

Page 5, figure 2
“PETs” is not defined

In panel (a)

There is no (a) label

In panel (b)

There is no (b) label

There is no explanation what the little boxes are.

Both ocean and atmosphere appear to happen at the same time. This is inconsistent with a sequential description.

Page 5, line 14
The surface boundary fields on the ocean surface is exchanged online

are

Page 6, line 4
but we updated it to couple

What is it

Page 6, line 12
In the present work, the Advanced Research WRF dynamic version (WRF-ARW, version 3.9.1.1) is used.

Include GitHub site for source code?

Page 6, line 32
In ESMF, ‘timestamp’ is a sequence of number,

numbers

Page 8, line 10

The time step for atmosphere simulation is 30 seconds.

For an approximately 9 km grid distance, 30 seconds seems overly conservative. Since WRF is the most expensive component, an increase to a 50 s timestep would be a substantial performance boost in overall timing. Is there a stability problem that is introduced with the coupling?

Page 8, line 21
net precipitation

Is this just accumulated precipitation minus evaporation? If so, just add a brief parenthetical.

Page 9, line 11-12

timescales of 10/0.5 days.

I am unfamiliar with what 10/0.5 days means.

Page 10, table 1

The four rows of the table should be more identifiable. There either needs to be more space between rows, or less vertical line space used in column three when the information extends to two lines.

The second ATM.STA should be ATM.DYN.

ERA5 is sufficient, without a description of bulk formula.

Page 10, line 10

validated against the ECMWF ERA5 dataset

Verifying a 9-km simulation with a 30-km reanalysis, specifically for cities that are along the coast may not be reasonable. It is not ever made clear how the max/min temperature comparisons are made. Does this field come out of ERA5? How is this information pulled from WRF?

Page 11, line 7-9

The simulation results obtained from coupled (CPL) run, the ERA5 data, and their associated difference are shown in Fig. 4 after 36 hours and 48 hours. It can be seen in Fig. 4(I) that the CPL run captures the heat wave event in the Red Sea region on June 2nd

If the simulated period started early (May 1, for example), is the June 2 heatwave event still present? Is this simply picked up because of the memory of the initial conditions? No where is it made clear what constitutes a heatwave event.

Page 11, line 19

all simulations can capture the T2 diurnal variation in the Red Sea region

Figure 4 shows that all simulations tend to have a larger T2 diurnal oscillation than the ERA5 reanalysis. This could be due to the cities are close enough to the coast that part of the 30-km grid cell contains moderating ocean temps, or that 30-km. Perhaps compare jobs to ERA5 (not necessarily to be shown in paper) to inform readers what is happening.

Page 12, figure 4

There is a systemic bias in ERA5: it is too cold at 1200 UCTC and too warm at 0000 UTC. Is this a good choice for validation?

There is little difference between the simulations. Most of the difference is between the simulated results vs ERA5. This is an indicator that this specific domain and this type of event may not be the best to showcase the capabilities of a coupled ocean atmosphere model.

Page 12, line 9

the SST in CPL run is tending to be similar to the realistic

This is not a clear way to state this point.

Page 12, line 13-14

It can be seen that four major heat waves (i.e., June 2nd, 10th, 17th, and 24th) and the T2 variations during the 30-day simulation are all captured

What is a heatwave event, and what defines captured?

Page 13, figure 5

This is more indication that this domain and case are not well chosen. Even after 23 days of constant SST, which would produce a poor representation of reality, the pattens displayed show virtually no difference over the land between CPL, ATM.STA, and ATM.DYN. Even with a bad SST, there is no discernible impact over land. A case needs to be chosen where a coupled ocean atmosphere model is relevant: flooding, precipitation, sea breezes, tides, inundation, typhoon cold wakes, steering flows for larger storms, etc.

Page 13, line 4

ERA5 uses a lower resolution grid and is unable to capture the T2 in the coastal city

Is this statement associated only with Yanuba, only with coastal cities, only when there is a large land-water temperature contrast? The authors should be careful about sweeping statements concerning ERA5.

Page 13, line 7-8

This may be due to the errors in initial conditions, or WRF physics schemes (e.g., land surface model, the PBL model) are unable to parameterize this extreme event.

A few lines up was “probably”, now we have “may”. These are model simulations, so you can determine cause. If the authors think the difference is due to initial conditions, back up the starting period by a few weeks. If the authors think the difference is due to physics, then try different combinations. Is there a specific assumption made in the chosen schemes? State a cause and defend the statement.

Page 13, line 9-10

In Fig. 6, the CPL run can better reproduce the evolution of the T2 compare to

ATM.STA run during the 30-day simulation:

Yes, CPL is better than ATM.STA, but that is a low bar. It is unphysical to have a constant SST for 30 days. Other than a single mention early in the paper concerning ATM.STA, the comparison should always be CPL vs ATM.DYN.

Page 13, line 12-14

We hypothesize that Mecca is much further away from the Red Sea than Yanbu and Jeddah, which indicates that the influence of air–sea coupling is strong near the coast.

First, Mecca is farther from the sea than the other two cities, that is not a hypothesis. What likely was intended was a hypothesis about the influence of air-sea coupling. OK, you have a stated conjecture, now conduct a test to support your position. Do you see diurnally varying on-shore/off-shore winds, for example? Is that signature missing farther inland? Is there flow that is blocked with a mountain range? If you remove the mountains, does the impact go further inland, etc.

Page 14, figure 6

It is difficult to identify the four heatwave events on the figure. Some background shading might be a good idea.

Choosing > 41 C in Jeddah seems appropriate, but you need > 45 C for Mecca and Yanbu. A heatwave event definition is required.

Is Yanbu too close to the water for the 30-km ERA5 reanalysis?

How are WRF and ERA5 daily max/min temperatures selected?

Page 15, figure 7

There is no statistical difference between the full coupled model and the ATM.STA (the experiment with a bad SST). More indication that this domain and case are insufficient to identify components of a functioning coupled system.

Page 15, figure 8

For the two top panels, some statistical information would be nice to allow the readers

to evaluate the CPL vs ATM.DYN vs ATM.STA tests.

Page 16, line 3-5

The daily SST fields from CPL run on June 2nd and 24th are shown in Fig. 9(I) and Fig. 9(VI). To validate the CPL run results, the SST fields obtained in OCN.DYN runs are shown in Fig. 9(II) and 9(VII) and the GHRSST fields are shown in Fig. 9(III) and 9(VIII).

Provide a brief explanation of what time was selected to verify with daily SST. Was the model SST averaged for a day? If there is no impact, just briefly state that.

Page 17, figure 10

Looking at the first week: explain (a) mean bias oscillation, and (a) RMS error ramp up.

Page 18, line 11

the air–sea interactions do not significantly impact the solar radiation

This comparison with observations also shows no impact with a coupled model.

Page 18, line 15-16

simulations over-estimated the total downward heat fluxes (CPL: 646 W/m^2 ; ATM.STA: 674 W/m^2 ; ATM.DYN: 663 W/m^2) for both heat wave events compared with MERRA-2 dataset (495 W/m^2)

This shows that the simulations (coupled vs non-coupled) are more similar to each other than to the verification. For this domain and case, this is more indication that the coupled model is not required. Or, if the uncoupled models are missing some important air-sea interactions, then the coupled model is not able to demonstrate the ability to capture those interactions either.

Page 18, 21-22

the present CPL simulations are capable of well capturing all the components of the surface heat fluxes during the heat wave events

You just showed that all cases are similar to each other, so all models are capturing components. While your statement is factually correct, it lends readers to believe that CPL is an improvement.

Page 18, line 28-30

On June 2nd, high-speed wind is observed in the northern and central Red Sea, and the CPL run successfully captures the small-scale features of wind speed patterns

From figure 14, all three shown tests (CPL, ATM.STA, ATM.DYN) are virtually identical, for both the June 2 and June 24 cases. This is disingenuous of the authors to imply that only CPL captures the small-scale features.

Page 18, line 31-32

the SST in the ATM.STA run is lower than the CPL run

Maybe it is more intuitive to say warmer or cooler when referring to temperatures.

Page18, line 34

The CPL run is able to capture

Figure 15 shows for the June 2 case, all tests perform well (since it is close to the SST initialization). For the June 24 case, the coupled model is similar to the ATM.DYN. This is and should be the important point that is made.

Page 19, figure 11 caption (same with page 20, figure 12 caption)

Only the heat fluxes over the sea is shown

are

Page 21, line 1-2

all simulation results are consistent on June 2nd because they are driven by the same initial condition

Not exactly, but after only a single simulated forecast day, the SSTs for all of the tests are similar.

Page 21, line 5-6

This is because the CPL run over-estimated the SST than the ATM.DYN run

Missing a word somewhere.

Page 22, line 6

runing 6-hour simulations

spelling

Page 22, line 8-9

When using 256 processors, there are 20480 cells (16 lat×16 lon×80 vertical levels) in each processor

Most of the grid points for the ocean are masked out, correct? Is this still an accurate statement of the amount of work on each MPI rank?

Page 22, line 9

there are 5120 overlap cells

What is an overlap cell? That term is not used in the atmosphere model. Is an overlap cell used in the ocean model or the coupler? How does it impact performance?

Page 22, line 13-14

It is noted in Fig. 16 that the parallel efficiency fluctuates when using 8 to 32 processors.

When describing the machine, include how many processes are used per node. Only comparisons with full nodes should be used for a good comprehension of scalability.

Page 22, line 14-15

This may be because of the fluctuation of the communication time, load imbalance, and I/O operations.

Here's that indefinite "may" again. To test I/O: turn off the output and only start timing after the input is complete. To test load imbalance, choose a domain that is not 84% land, etc. To test fluctuating communication time, are the timing results reproducible. There is no need to be uncertain.

Page 23, line 4-5

The atmospheric model also uses a smaller time step (30 s) than that of the ocean model (120 s) and has more complex physics parameterization packages

Within the atmosphere model, you could legitimately compare the complexity of one scheme vs another, but probably not between the ocean model vs the atmosphere model. The time step comparison, the relative number of solved grid cells, and the cost per grid cell from ATM.DYN vs OCN.DYN would be sufficient.

Page 23, line 9

We hypothesis that the cost of the ESMF/NUOPC coupler is communication cost and it becomes important as the amount of computation work is reduced with the number of grid cells in these strong scaling tests.

We hypothesize

This is a poor attempt to describe what is happening with strong scaling.

Also, "the cost … is communication cost" should be cleaned up.

Page 24, figure 16 caption

The simulation with the smallest case is regarded as base case when computing the speed-up

What does the smallest case mean? For strong scaling, the cases are the same size.

Page 25, line 6

The development activities has been focused on providing

have

Page 25, line 14-15

Improvements of the coupled model over the stand-alone simulation with static SST forcing are observed in capturing the T2, heat fluxes, evaporation, and wind speed.

This should focus on CPL vs ATM.DYN, not CPL vs ATM.STA. Also, quite a number of the authors' validations showed that even CPL vs ATM.STA was reasonable. This points to a problem in the fundamentals of the setup. If a month-long fixed SST is giving indistinguishable results to the new coupled model, at best the coupled model is doing no worse than the poorly constructed comparison case. An appropriate domain and case need to be selected.

Page 25, line 17-18

The coupled model scales linearly for up to 128 CPUs and the parallel efficiency remains about 70% for 256 processors.

This is a meaningless statement without a total number of horizontal grid points included, without the relative number of ocean vs atmosphere grid cells, without the cost per grid cell in each model, etc.

Page 25, line 19-20

Hence the coupled model can be applied for high-resolution coupled regional modeling studies on massively parallel processing supercomputers.

High-resolution has nothing to do with the number of grid cells. Arguably, 9 km is not

actually high-resolution.

No one will associate 256 MPI processes with massively parallel.

Page 25, line 22-23

These preliminary results motivate further studies in evaluating and improving this new regional coupled ocean–atmosphere model for investigating dynamical processes and forecasting applications in regions around the globe where ocean–atmosphere coupling is important.

That is what *this* paper should be about. There is no way to evaluate this new coupled model if the model is not used over a domain and for a case that would require coupled ocean and atmosphere capabilities.