Reviewer comments
Revised manuscript
Replies to the reviewers

Reviewer 1:
While a heatwave event is high impact, it is not the first general episodic type that comes to mind when specifically highlighting the capabilities of a coupled ocean-atmosphere model.

Reply: We thank the reviewer for the insightful comments. We agree with the reviewer that the simulation of the heatwave events is not an ideal case for testing a regional coupled model and its benefits. Yet our goal for this work and manuscript is to (1) introduce the design of a newly developed regional coupled ocean-atmosphere modeling system with a state-of-the-art coupler (ESMF with NUOPC), (2) describe the implementation of the modern coupling framework, (3) validate the coupled model using a real-world example, and (4) demonstrate and discuss the parallelization of the coupled model. This work is the first time WRF has been coupled to the MITgcm, and we are using this manuscript to present the new resource. We now highlight the repository for downloading this resource in the abstract. We chose the location because our funding from KAUST was obtained to develop and implement the model for the Red Sea region. We chose the heatwave event due to their societal importance. This test case is meant only as a proof-of-concept example to show that the regional coupled model is working as one would expect, even in an extreme weather example. We have edited the abstract and introduction of the manuscript to emphasize our motivation and goal further and to make it clear that our objective is not to test the forecast skill of the model in this work. We are doing follow-on work to test the model skill in various scenarios. Our focus here is why we have chosen to submit this manuscript to Geoscientific Model Development, which promotes publication of code development and technical aspects of geoscience research.

Demonstrating the impact of coupled models is a research topic that requires a significant effort into the future. We have tried to make the text clear that our focus here is solely on the technical development of a coupled ocean-atmosphere model, implemented using state-of-the-art components. We hope this addresses the concerns of the reviewer about the test case satisfactorily.

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.

Reply: Yes, the ERA5 has coarser resolution and the strong gradient from water to desert has been more carefully handled. The ERA5 reanalysis uses a fractional value between 0 and 1 as the land-sea mask. In the old manuscript, we used bilinear interpolation and under-estimated the daily-high temperature in Yanbu. We now consider the effect of the land-sea mask and updated our figures.

Regarding performance, we have removed all language implying our coupled model is superior to ERA5. In keeping with the manuscript focus, the language now reflects the fact that the coupled system we developed is performing as one would expect.

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.

Reply: Because we aim to report on the technical development of a new coupled model, we compare CPL, ATM.STA and ATM.DYN runs to show the SST and atmospheric boundary variables are reasonably updated in the coupled model. We have revised the manuscript to focus on the code development and technical aspects. In our work, we emulate previous studies which implemented regional coupled models with other components and use a stationary SST for testing the implementation of the model [Loglisci et al. 2004, Warner et al. 2010]. Differences from ATM.STA gives a measure of how fast the SST is evolving and the size of the evolved SST signal the coupled model is producing.

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.

Reply: We have edited the manuscript and replaced these terms by directly attributing the uncertainty in these statements. We have removed the other uncertain statements that we are not able to test by using our model.

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.

Reply: We agree with the reviewer that this was not clear in our previous manuscript. There is no universal definition for heat waves, as this is a definition that varies from region to region. Hence, we have replaced the term 'heat waves' with 'heat events' in the manuscript as we are trying to simulate the events that had record high temperature in this region.

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.

Reply: We agree that ATM.STA and ATM.DYN tests also perform well. We aim to demonstrate the coupled model is capable of performing coupled simulations. The model tests are used to validate the coupled model is performing the coupled simulation as expected. We have revised the manuscript to emphasize our purpose.

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

Reply: We are grateful that the reviewers mentioned these incomplete comments. We have revised our manuscript and have gone through the scaling test to make sure we finished all our comments.

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.

Reply: We have revised the labels of Figs. 1 and 2. We have also gone through the figure captions to improve them.

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

Reply: We thank the reviewer again for this. We have proofread the manuscript.

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.

Reply: We select the case to illustrate our code development and validate our implementations on technical aspects, but not to provide an in-depth analysis of ocean-atmosphere coupling in the region. We have explained the choice to model this region and the extreme heat events in the revised manuscript:

We simulate a series of heat events in the Red Sea region, with a focus on validating and assessing the technical aspects of the coupled model. There is a desire for improved and extended forecasts in this region, and future work will investigate whether a coupled framework can advance this goal. The extreme heat events are chosen as a test case due to their societal importance. While these events and the analysis here doesn't highlight the value of coupled forecasting, these real-world events are adequate to demonstrate the performance and physical realism of the coupled model code implementation.

Specific Comments
Page 4 line 2:
is shown in Fig. 1(a)
There is no (a) label

Reply: We have now fixed this typo.

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.

Reply: PETs are Persistent Executions Threads, which are single processing units (e.g., CPU, GPU) defined by ESMF. We defined them in the manuscript but not under the caption of Fig. 2. Now it is added. We have now added (a) and (b) labels to Fig. 2. In panel (b), each little box is the decomposed domain. We have added the explanation in the captions.

In Fig. 2(b), in the sequential mode, each PET works in a sequential way. The small arrows in the timeline show the sequential nature of the implementations. The coupled model first integrates the ocean component, then integrates the atmosphere component. After the integration is finished in one coupling step, the ESMF connector gathers the outputs from ocean/atmosphere components and update the boundary condition for the next step. We also implemented the concurrent mode and added it in Fig. 2(b):

Panel (b) and (c) shows the sequential and concurrent mode implemented in SKRIPS, respectively. PETs (Persistent Execution Threads) are single processing units (i.e., CPU, GPU) defined by ESMF. OCN, ATM, and CON denote oceanic component, atmospheric component and connector component, respectively. The blocks under PETs are the CPU cores in the simulation; the small blocks under OCN or ATM are the small sub-domains in each core; the block under CON is the coupler. The red arrows indicate the model components are sending data to the connector and the yellow arrows indicate the model components are reading data from the connector. The horizontal arrows indicate the time axis of each component and the ticks on the time axis indicate the coupling time step.

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

Reply: We have now fixed this typo.

Page 6, line 4
but we updated it to couple
What is it

Reply: We now replaced 'it' using 'the baseline coupler'.

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?

Reply: We now have added the Github site for WRF:

WRF is used extensively for operational forecasts (http://www.wrf-model.org/plots/wrfrealtime.php) as well as realistic and idealized dynamical studies. The WRF code and documentation are under continuous development on Github (https://github.com/wrf-model/WRF).

Page 6, line 32
In ESMF, 'timestamp' is a sequence of number,
numbers
Page 8, line 10

Reply: We have now fixed this typo.

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?

Reply: We agree with the reviewer that increasing the time step can improve computational performance. However, the computational domain is complex in the Ethiopian Highlands and Hejaz Mountains (Saudi Arabia). [Hughal et al. 2017; WRF forum] We find that the CFL number is larger than 1 if we use a 50 second timestep. We have revised the manuscript:

The time step for atmosphere simulation is 30 seconds, which is chosen to avoid violation of the CFL condition.

[Hughal et al. 2017. Wind modelling, validation and sensitivity study using Weather Research and Forecasting model in complex terrain. Environmental modeling & software]
[http://forum.wrfforum.com/viewtopic.php?f=8&t=357]

Page 8, line 21
net precipitation
Is this just accumulated precipitation minus evaporation? If so, just add a brief parenthetical.

Reply: The MITgcm gets precipitation from WRF and uses the precipitation to calculate the freshwater flux (E-P). We have replaced 'net precipitation' with 'precipitation'.

Page 9, line 11-12
timescales of 10/0.5 days.
I am unfamiliar with what 10/0.5 days means.

Reply: What we meant to say is 'the inner and outer boundary relaxation timescales of the sponge layer are 10 and 0.5 days, respectively'. We have rewritten this sentence.

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.

Reply: We have revised the tables and made the four rows more identifiable. We agree that ERA5 is sufficient and removed the description of bulk formula. We also fixed the typos.

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?

Reply: We agree with the reviewer that verifying the surface air temperature using the 30-km ERA5 data may not be reasonable. Hence we used the NCDC ground observations to validate the air temperature. The T2 fields are interpolated to the NCDC stations for both WRF and ERA5. The NCDC station data are the best observations available in this region. Unfortunately the observation time in NCDC data is not available, hence the daily max/min T2 every 24 hours from the model simulations are compared with the daily max/min T2 from the NCDC data. We have added a discussion on this in the revised manuscript:

Since the 30-km resolution ERA5 dataset may not be adequate to capture the sharp T2 gradients near the coast, the T2 is also compared with the available ground observations from NOAA National Climate Data Center (NCDC climate data online at http://cdo.ncdc.noaa.gov/CDO/georegion). To evaluate the modeling of T2 in three major cities near the eastern shore of the Red Sea, the T2 fields from both the simulations and ERA5 are interpolated to the NCDC stations. The daily maximum and minimum temperatures are compared with those from NCDC data.

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.

Reply: This is a test validation case to show that the coupled model works as expected. We agree that it is not proper to claim that the CPL run captures the heat events without a detailed discussion on what constitutes the events. We have revised our manuscript to simply emphasize that the performance is reasonable for this high heat events.

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.

Reply: We agree with the reviewer that the simulations have larger T2 diurnal oscillation than ERA5 and we have rewritten this sentence:

Fig. 4 also shows the diurnal variation of T2 in the Red Sea region, and the diurnal variation will be further discussed later in this section.

We also agree that because the cities are on the sea we need to account for the land masks when interpolating. We have updated Fig. 6 with a better accounting for this mask, and the daily max/min T2 in ERA5 now agrees better with the NCDC observations.

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?

Reply: We agree with the reviewer that ERA5 is colder at 1200 UTC and warmer at 0000 UTC on land compared with the simulations. The daily max/min T2 in ERA5 agrees better with the NCDC observations compared to all simulations. Hence we use ERA5 to validate the T2 fields obtained in the simulations with the aim being to demonstrate that the coupled model is capable of reasonable simulations of the coupled ocean-atmosphere system. We have revised our discussion in the manuscript:

Since ERA5 air temperatures are in good agreement with the NCDC ground observations in the Red Sea region (detailed comparison of all stations are not shown), we use ERA5 data to validate the simulation results……. To validate the coupled ocean-atmosphere model, the mean T2 differences over the sea in the simulations are compared with the ERA5 data. The mean T2 biases and RMSEs over the sea are shown in Table 3. The biases of the T2 are comparable with the biases reported in other WRF simulations for heat events [Imran et al, 2018].

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.

Reply: We agree with the reviewer that the differences in the simulations are subtle. Fig. 4 shows our model simulation does not have much error compared with ERA5 and benchmark WRF simulations (the bias and RMSE of T2 in the present work are similar to those in the benchmark WRF-ARW simulations [Xu et al. 2009, Zhang et al. 2013, Imran et al. 2018]). In Fig5, T2 differs very little on land, but the T2 in the CPL run is about 1.0 degC warmer than the ATM.STA run and has smaller error than ATM.STA run. This shows the coupled model driven by a warming SST can better capture the surface T2. We also show that the coupled run can be as good as uncoupled run using an updated SST. We have revised our manuscript:

It can be seen in the figure that the T2 errors on land are consistent for all three simulations. However, the T2 over the sea in CPL simulation has smaller mean biases with the validation ERA5 data (10th: -1.24 degC; 24th: -0.81 degC) compared with the ATM.STA run (10th: -1.56 degC; 24th: -1.83 degC)…… The T2 over the water in the CPL run is closer to the ERA5 because MITgcm in the coupled model provides updated warming SST, which warms the T2; the ATM.STA run uses a constant cooler SST from June 1st, and thus the T2 is determined by the constant cooler SST. On the other hand, when comparing the CPL run with the ATM.DYN run, the mean difference is smaller (10th: +0.04 degC; 24th: -0.62 degC). This shows the CPL run is comparable to the ATM.DYN run driven by an updated warming SST.

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.

Reply: We agree with the reviewer that our statement is not clear in the old manuscript. We have removed this sentence in the manuscript.

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?

Reply: Since there is no universal definition on the heat waves, we have replaced heat waves using heat events in the manuscript. We have also revised 'capture the heat wave events' by showing the difference of the maximum daily temperature between difference simulations.

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 patterns 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.

Reply: We agree with the reviewer that after 23 days of constant SST, the patterns on land does not show significant differences. However, the ocean T2 differs by about 1 degC in Fig. 5(IX), showing the impact of SST. In our paper, we focus on the technical issues and the development of the coupled

model and validate that is performing in a reasonable way. This is why we have chosen to submit this manuscript to Geoscientific Model Development, which considers the publication of code development and technical aspects of research.

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

Reply: The authors thank the reviewer for pointing out this. We agree that our statement concerning ERA5 in this sentence is inaccurate and too general. We have revised our interpolation method near Yanbu by adding the land-sea mask. We have removed this sentence in the revised manuscript.

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 the 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.

Reply: Yes, we have tried a few different WRF physics options and different combinations (e.g., YSU/MYJ/MYNN + Kain-Fritsch/Zhang-McFarline + Noah LSM/Noah-MP LSM/RUC + MM5/MYNN). We tried different combinations to help us model the diurnal cycle. In our test, only Kain-Fritsch+RUC can capture the diurnal cycles in the Red Sea region. Other combinations will overestimate the daily max T2 and underestimate the daily min T2 on land. Since our manuscript focuses on the development of the model, we have removed the discussion of the errors in the manuscript.

Page 13, line 9-10

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

Reply: We agree with the reviewer that it is unphysical to have a constant SST for 30 days. But we are using both ATM.STA and ATM.DYN cases to validate the coupled model, as done in previous studies [Loglisci et al. 2004, Warner et al. 2010]. However, since the difference is insignificant we have removed this discussion.

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.

Reply: The authors thank the reviewer for pointing out this. There is no mountain between Mecca and the sea. We aim to present the development of the coupled model and we have removed the discussion of this in our manuscript.

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.

Reply: The authors thank the reviewer for pointing out this. There is no universal definition of heat wave events. We have replaced 'heat wave' using 'heat events' in the manuscript.

Is Yanbu too close to the water for the 30-km ERA5 reanalysis? How are WRF and ERA5 daily max/min temperatures selected?

Reply: Yes, the ERA5 resolution may not be able to resolve the strong water to desert transition. The ERA5 reanalysis uses a fractional value between 0 and 1 as the land-sea mask. In the old manuscript, we used bilinear interpolation and under-estimated the daily-high temperature in Yanbu. We now consider the effect of the land-sea mask and updated our figures. The WRF/ERA5 daily max/min temperatures are the max/min values every 24 hours. We give the method specifics in the revised manuscript.

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.

Reply: The coupled model better forecasts the daily low temperature in Jeddah and Yanbu by about 1.0 and 0.5 degC from day 20 to day 30, and we added the discussion on this. We agree with the reviewer that showing the mean value does not show any significant improvement and we plotted the RMSE values instead in this figure. We have added the statistic information in Figure 6.

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.

Reply: We have added the statistical information in Fig. 8. We have also added the information for other figures.

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.

Reply: GHRSST uses the nighttime SST and we also use the nighttime SST at 0000 UTC (about 3 A.M local time) to compare with the GHRSST. The snapshots of the SST obtained in the CPL run are also compared with the available HYCOM data.

Page 17, figure 10
Looking at the first week: explain (a) mean bias oscillation, and (a) RMS error ramp up.

Reply: We performed new simulations using 3-hourly HYCOM boundary conditions. When using this 3-hourly data the mean bias oscillation is not observed in the first week. The initial SST error is zero because both simulations are initialized using HYCOM in Fig. 10(a), but the HYCOM dataset is corrected by observations so that the error increases. We have added the discussion on RMSE increase in the manuscript.

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.

Reply: We agree with the reviewer that the impact of the coupled model is small. We have added this in our manuscript.

Page 18, line 15-16
simulations over-estimated the total downward heat fluxes (CPL: 646 W/m2 ; ATM.STA: 674 W/m2 ; ATM.DYN: 663 W/m2) for both heat wave events compared with MERRA-2 dataset (495 W/m2)
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.

Reply: We have attempted to better highlight differences. We have removed the snapshot comparisons (previous Fig. 12) and instead to better quantify the heat fluxes we calculate the time series of the mean deviation and RMSE between the model simulations and MERRA-II.

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 leads readers to believe that CPL is an improvement.

Reply: We agree with the reviewer that our discussion on the surface heat flux ignored the uncoupled simulations. We have re-written this in the revised manuscript:

It can be seen in the figure that both the CPL and ATM.STA runs reproduce the mean heat flux over the Red Sea estimated by MERRA-2.

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.

Reply: We agree with the reviewer that our discussion on U10 ignored the uncoupled simulations. We have re-written this in the manuscript:

both the CPL and ATM.STA runs successfully capture the small-scale features of wind speed patterns

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.

Reply: We have replaced 'lower' using 'cooler' in this sentence. We have also replaced higher/lower using warmer/cooler when describing SST.

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.

Reply: We agree with the reviewer that we did not make the important points clear in the discussion of the evaporation. Because the evaporation is proportional to the latent heat flux, we have moved it to the appendix. But, we still revised the discussion on the evaporation.

Page 19, figure 11 caption (same with page 20, figure 12 caption)
Only the heat fluxes over the sea is shown
are

Reply: We have fixed this typo.

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.

Reply: We agree with the reviewer. We have replaced the old sentence with:

After 48-hours, the simulation results are close with each other (e.g., the RMSE between CPL and ATM.STA simulation is smaller than 10 cm/year).

Page 21, line 5-6
This is because the CPL run over-estimated the SST than the ATM.DYN run
Missing a word somewhere.

Reply: We replaced the old sentence with 'This is because the SST in the CPL run is warmer than the ATM.DYN run'.

Page 22, line 6
runing 6-hour simulations
spelling

Reply: We have now fixed this typo.

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?

Reply: We agree most of the ocean points are masked out. We have rewritten this sentence in the original manuscript:

When using 256 CPU cores, there are a maximum of 20480 cells (16 lat x 16 lon x 80 vertical levels) in each core. It is noted that the ocean model only solves the Red Sea (16% of the domain) and most of the ocean points are masked out in this real-world test.

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?

Reply: The 'overlap cell' is not an appropriate expression. We have removed this sentence in our manuscript and rewritten it:

When using 256 CPU cores, there are a maximum of 20480 cells (16 lat x 16 lon x 80 vertical levels) in each core. It is noted that the ocean model only solves the Red Sea (16% of the domain) and most of the ocean points are masked out in this real-world test. From results reported in the literature, the parallel efficiency of the coupled model is comparable to other ocean-alone or atmosphere-alone models when having similar number of grid points per CPU core [Marshall 1997, Zhang 2013].

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.

Reply: We agree with the reviewer. We performed 'full node' test using another cluster (because our old cluster is down). We have added:

We started using N_p0 = 32 because each compute node has 32 CPU cores.

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.

Reply: We agree with the reviewer. We did perform a cleaner test by turning off the unnecessary I/O. Nevertheless, the purpose of Section 5 is to show the coupler has good parallel efficiency and does not slow down the simulation. We have removed the discussion on the fluctuation of the parallel efficiency.

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.

Reply: We have now removed the discussion on the parameterization schemes in our manuscript:

The atmospheric model is much more time-consuming because it solves the entire computational domain, while the ocean model only solves the Red Sea (16% of the domain). The atmospheric model also uses a smaller time step (30~s) than that of the ocean model (120~s). If a purely marine region is selected in an ideal case, the cost of ocean and atmosphere models would be more equal compared with this Red Sea case.

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.

Reply: We thank the reviewer for pointing out this. We turned off unnecessary I/O in a clean test. The CPU time spent on the atmosphere model, ocean model, and the coupler are reported in the CPL run. We have replaced the old sentences with:

The coupling process takes less than 3% of the total costs in the CPL run. Although the proportion of the coupling process in the total costs will increase when using more CPU cores, the total time spent on the coupling process is similar. The CPU time spent on two uncoupled runs (i.e., ATM.STA, OCN.DYN) is also shown in Table. 3. Compared with the uncoupled simulations, the ESMF-MITgcm and ESMF-WRF interfaces do not increase the CPU time in the coupled simulation. In summary, the scalability test results suggest that the ESMF/NUOPC coupler is not a bottleneck for using SKRIPS in coupled regional modeling studies.

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.

Reply: We have replaced 'the smallest case' using 'the simulation with 32 CPU cores'.

Page 25, line 6
The development activities has been focused on providing
have

Reply: We have fixed this typo.

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.

Reply: The purpose of this paper is to present the coupled modeling system and not to answer scientific questions on the impact of coupling. We have removed the discussion on the improvement of the coupled model in the conclusion.

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.

Reply: We thank the reviewer for pointing out this. We have rewritten this sentence:

The parallel efficiency of the coupled model is consistent with that of the stand-alone ocean and atmosphere models when using various numbers of CPU cores in the test.

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.

Reply: We agree with the reviewer that high-resolution and massively parallel are not used appropriately in the conclusion. We have removed 'high-resolution' and 'massively parallel' in the conclusion section and rewritten this sentence:

The CPU time associated with different parts of the coupled simulations is also presented, suggesting the ESMF/NUOPC driver is not the bottleneck of the computation. Hence the coupled model can be implemented for coupled regional modeling studies on supercomputers with comparable performance as that attained by uncoupled stand-alone models.

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.

Reply: We agree with the reviewer that the heat events do not show the strong capability of a coupled model. However, our aim of this paper is to detail the development of this coupled model and validate the performance. Investigating dynamical processes is outside the scope of this manuscript. We have re-written the abstract and introduction of the manuscript to emphasize our motivation and goal further and make it clear that our objective is not to test the forecast skill of the model in this work,

but instead to present this new coupling model resource. We are doing follow-on work to test the model skill in various regions and scenarios. This is also why we submit this manuscript to Geoscientific Model Development, which considers the publication of code development and technical aspects.

Reviewer 2:
The authors of the manuscript coupled an atmospheric model with an oceanic model using ESMF infrastructure. They evaluate the model performance during June 2012, which contained three heat waves over the Arabian Peninsula and the Red Sea. The model simulation lasted one month. The skill was better during the earlier simulation timeframe.

Major comments
The paper still needs focusing. I am still struggling to see the main point of the paper: was it a technical challenge that had to be overcome, or was it a science question that could finally be answered with the coupled system? Is the system designed for seasonal prediction, hence a 30-day run? If it is focused on the heat waves, why is the bulk of evaluation over the Red Sea? When resolving a phenomenon that goes through a diurnal cycle, such as air temperature at 2 meter height, isn't it imperative that the boundary conditions can resolve the diurnal cycle? Obtaining daily HYCOM values completely misses the diurnal variability, and SST change can be up to a few degrees Celsius, affecting surface heat fluxes.

Reply: We aim to describe the development of coupled code and use an example case to show the coupled system works. We are introducing a new resource, and we are not trying to answer a specific science question in this paper. We have added a discussion on the aim of this paper in the abstract and the introduction. We aim to validate the coupled ocean-atmosphere model and we demonstrated the capability of this model by showing the surface temperature, SST, heat flux, and wind speed are captured over the ocean compared with validation data. We now present the repository where one may download this resource in the introduction.

We agree that using a 3-hourly boundary condition will better resolve the diurnal cycle. We have replaced the daily boundary condition with 3-hourly values to capture the diurnal SST cycle of HYCOM in the region influenced by this boundary condition.

Please re-read the manuscript carefully before submitting it. The second version still contains typos.

Reply: Thanks. We have gone through the manuscript carefully this time and fixed the typos.

Minor comments
P1 l12: I am not sure if the coupled system has good skill. Some errors are quite big.

Reply: We aim to describe the development of coupled code and use an example case to show the coupled system works. We agree with the reviewer that the presented test does not show the coupled model has *better* skill than uncoupled simulations. We have revised our manuscript to reflect this.

We also agree that some errors in the heat fluxes at daytime are quite large at a few snapshots. But (1) the time-averaged error is much smaller, and (2) our model error is comparable with the errors reported in the literature by using different WRF radiation schemes (Zempila et al. 2016). We also

tried different radiation schemes and our conclusions are consistent with the literature. We now plot the time-series of the heat fluxes and have moved the snapshot evaluations to the appendix.

P7 l12: The sequential mode might be simpler to start, but the overall performance of the system in your case will suffer – the ocean component has fewer active (water) points on which the computation is performed, thus affecting load balancing. If discussing scalability of the system, this is not the best choice.

Reply: We have now implemented both concurrent and sequential modes. We agree that the sequential mode suffers and slows down the parallel efficiency. We have revised our manuscript:

The ESMF allows the PETs to run in sequential mode, concurrent mode, or mixed mode (for more than three components). We implemented both the sequential and concurrent mode in SKRIPS, shown in Fig. 2(b). In sequential mode, a set of ESMF gridded/coupler components run in sequence on the same set of PETs. At each coupling time step, the oceanic component is executed when the atmosphere component is completed or vice versa. On the other hand, in concurrent mode, the gridded components are created and run on mutually exclusive sets of PETs.

P8 l16: If I understand correctly MITgcm also runs on 256x256 grid, of which a large portion are idle (land points)?

Reply: Yes. A large portion of the MITgcm ocean model is idle.

P9 l9: The link to HYCOM data is incorrect – it should read '/dataserver/'.

Reply: We have now fixed this typo.

P10 Table1: The typo has not been corrected; it still reads ATM.STA instead of ATM.DYN.

Reply: We have now fixed this typo.

P11 l7: For clarity: 'from the model experiments', probably better than 'from various experiments'.

Reply: We have replaced the original manuscript using 'from the model experiments.

P11 l10: I am not convinced ERA5 air temperature is in good agreement with NCDC ground observations. Figure 6 indicates discrepancies up to 10 degrees Celsius!

Reply: We thank the reviewer for pointing out this. The difference is 10 degrees because we did not consider the land-sea mask when interpolating the ERA5 air temperature. We have considered the land-sea mask and re-done the interpolations in the revised manuscript.

P11 l15: Why are T2 differences discussed for hours 36 and 48?

Reply: First, the extreme heat event in Mecca is observed on June 2nd, which is 36 hours after the initialization. The T2 after 48 hours are presented to show the diurnal variation of T2 in the Red Sea. In addition, we present the T2 differences after 36 and 48 hours to show the coupled model is consistent with the uncoupled models when the SST are close in the simulations. We have added the discussion in Section 4.1.

P12 Figure 4: The model is too hot over land at daytime (36 hours) and too cold almost everywhere at night (48 hours). The temperature differences are significant!

Reply: We agree that the temperature differs by a few degrees in our simulations. We tried different combinations of WRF schemes and other combinations have much larger errors in T2 compared with the present combinations. We focused on validating the coupled model with respect to the uncoupled atmospheric run give prescribed SST, and we now emphasize this in our manuscript.

P12 l5: The CPL run results are closer to the ERA5 dataset where? Over land, over water?

Reply: The CPL run results are closer to ERA5 over water. We have revised this in the manuscript.

P12 l8: If SST fields from CPL and ATM.DYN are similar, it does not mean 'SST in CPL run is tending to be more similar to the realistic'. Please reword.

Reply: We have removed this sentence.

P12 l13: NCDC has already been defined.
Reply: We have removed the redefinition of NCDC.

P12 l16: Ground observations and ERA5 are not in good agreement according to Figure 6.

Reply: The difference is because we did not consider the land-sea mask when interpolating the ERA5 air temperature. We now consider the land-sea mask and have re-done the interpolations. Now the NCDC observation and ERA5 are consistent.

P13 Figure 5: The surface air temperature is persists to be too warm over land at daytime seen at t=9.5 days and t=23.5 days). Care to comment?

Reply: We agree that the temperature differs by a few degrees in our simulations. We hypothesize this is because WRF land surface schemes do not capture T2 in the desserts in the simulations, but we need detailed numerical tests to defend our claims. Since our aim of this paper is to show the technical development of a new coupled model, we did not investigate the T2 bias on land.

P13 l3: What is RMSE difference?
Reply: Yes, it is RMSE, not RMSE difference. We have removed the 'difference' in the manuscript.

P13 l5: Are you implying that the simulation was not capable of capturing the June 2nd event was not captured in Mecca? I do not understand the meaning of the sentence.

Reply: We want to say the daily maximum T2 is not captured well by the models in Mecca, but it is captured in Jeddah and Yanbu. We have replaced the old sentence:

It should be mentioned that both the present simulations and ERA5 reported a T2 that is 2.8 degC cooler than the observed record-high T2 in Mecca on June 2nd. This under-estimation is comparable with the RMSE of the daily high T2 in Mecca (2.25 degC in CPL run).

P13 l7: If initial conditions were at fault, the first day maxima would not be captured. Why is the WRF physics scheme in error – it worked on day 1.

Reply: We agree with the reviewer that it is not because of the initial condition. We have removed it in our manuscript.

P13 l8: Physics schemes do not parameterize extreme events. They parameterize physical processes that can not be explicitly resolved in a numerical model.

Reply: We agree with the reviewer. We hypothesize that the physical parameterization schemes used are problematic for extreme events. However, since this paper focuses on the development of the coupled model, we did not check the detailed implementation of the WRF schemes. We have removed this sentence in our manuscript.

P13 l10: …'T2 compare'… should read 'T2 compared'.

Reply: We have now fixed this typo.

P14 Figure 6: Can you mark the heat event episodes with light gray background?

Reply: We have highlighted the events that we looked at in Fig. 6.

P14 Figure 6: The discrepancy between ERA5 and observed maxima and minima is troublesome if ERA5 is used to evaluate the model results.

Reply: The difference is because we did not consider the land-sea mask when interpolating the ERA5 air temperature. We now consider the land-sea mask and have re-done the interpolations. After the land-sea mask is considered, the NCDC observation and ERA5 are consistent.

P15 Figure 8: 'air temperature' should read 'air temperature over the Red Sea'. I do not understand why only over water.

Reply: We have replaced 'air temperature' using 'air temperature over the Red Sea' in the caption. We don't use the land surface temperature because we want to investigate the performance of the coupled ocean-atmosphere model.

P15 Figure 8: Can you comment on the bias and rmse time drift of the ATM.STA results?

Reply: Yes. We have added:

The bias and RMSE of T2 in the present work are similar to those in the benchmark WRF-ARW simulations [xu et al. 2009, zhang et al. 2013]. The differences of the mean bias and RMSE between the simulations and ERA5 data are also plotted to demonstrate the evolution of the CPL errors compared with ATM.STA and ATM.DYN runs. It can be seen that the CPL run has smaller bias and RMSE than the ATM.STA run throughout the entire simulation.

P16 l6: It appears that the model captures the general meridional SST gradient in the Red Sea. But the actual temperature differences are quite large (Figure 9).

Reply: We have re-done the simulations and used 3-hourly HYCOM data as the boundary condition. The actual temperature difference is much smaller when using 3-hourly data.

P16 l14: I am afraid that daily data is not frequent enough to study diurnal phenomena. I am not familiar with HYCOM data on the aforementioned server, but I would be surprised if a global ocean model provides fields only once a day.

Reply: We now run the simulations using 3-hourly HYCOM data. The coupled model can capture the diurnal cycle better when driven by the 3-hourly boundary condition.

P17 l5: 'This is because the HYCOM data is cooler than GHRSST'. Have you considered the difference in HYCOM bulk SST at the topmost model level, and the actual skin SST measured by satellites, reported by GHRSST? Also, the temperatures might be lower, but they don't make the data cooler.

Reply: We have clarified this statement in our manuscript.

This is because our models are initialized by using HYCOM/NCODA, and the temperature in the topmost model level is cooler than the estimated foundation SST reported by GHRSST.

P17 l11: Why are surface heat fluxes evaluated only over the Red Sea?

Reply: We focus on the Red Sea because we are validating our coupled model. We have added the explanation of this in the manuscript.

P18 l3: CPL and ATM.DYN runs exhibit more latent heat fluxes coming out of the ocean. Why? And reword please.

Reply: We have revised this paragraph by discussing the time series of the latent heat fluxes. The previously discussed snapshots may not be adequate to describe the capability of the coupled model. When comparing the hourly latent heat fluxes, the simulations do not have significantly more latent heat flux compared with the MERRA-II data.

P18 l6: Latent heat, or latent heat fluxes (three times).

Reply: We have now fixed the typos.

P18 l8: Why is latent heat flux adequate close to the coastline?

Reply: We have revised this. The MERRA-2 data have coarsened resolution and coastal region is 'contaminated' by the land points [Kara et al. 2008, Gelaro et al. 2017]. We have ignored the partial land points in the interpolation and re-calculated the heat fluxes.

P18 l10: I am not convinced the short and longwave radiation fluxes are captured reasonably well. According to Figure 12, the differences are almost up to 30% of the total forcing.
P18 l15: Yes, the differences are almost 200 W/m2!
P18 l17: Points with discrepancies can be masked out to help with the evaluation.
P18 l21: I do not agree with the conclusion that 'overall the CPL simulations are capable of well capturing all the components of the surface heat fluxes'.

Reply: We have removed Fig. 12 from the new manuscript. We agree with the reviewer and have concluded that a snapshot example is inadequate to show the difference of the heat fluxes. Instead, we now calculate the time series of the surface heat flux and show the mean deviation and RMSE between the coupled model and MERRA-II. The RMSE is 105 Wm-2 and higher if one only considers

daytime. We compare our heat flux simulation with the literature [Zempila et al. 2018, Imran et al. 2018]; the uncertainty in the hourly flux can be 50% to 100% for a few widely used WRF schemes. We have also revised our manuscript.

P18 l24: How is latent heat flux different from evaporation?

Reply: We agree that the evaporation is a function of the latent heat flux in the coupled model. Hence, we have moved the evaporation part to the appendix.

P18 l24: Surface winds are winds at 10 meters? Are there any wind observations in the NCDC data? Why not compare sea level pressure? If the pressure field is not correct, that could explain the discrepancy in winds.

Reply: Yes, surface winds are winds at 10 meters and we now mention this in the revised manuscript. Unfortunately, the U10 data is not available from the NCDC data. Yes, we agree that presenting the sea level pressure is also helpful for investigating the difference in surface wind fields. However, here we solely present U10 to illustrate that the coupled model is capable of simulating the surface momentum transfer. We have added this in our revised manuscript.

P18 l32: Please reword: 'is small than'.

Reply: We have now fixed this typo.

P21 l5: Please reword: 'This is because the CPL run over-estimated the SST than the ATM.DYN run'.

Reply: We revised that sentence with 'This is because the SST in CPL run is warmer than that in the ATM.DYN run'.

P21 l7: Did the model create any precipitation? If there is none observed is not a sufficient reason not to include it in the discussion.

Reply: The model observed precipitation in the southern Red Sea and the Ethiopian Highlands. But the precipitation is not observed in the three major cities. We agree with the reviewer and we have removed the discussion on the precipitation.

P22 l9: What are overlap cells? Why does it matter they are 25% of the total cells?

Reply: Section 5 aims to show the coupler is not the bottleneck. Hence, we have removed the discussion of 'overlap cells' and rewritten this paragraph:

When using 256 CPU cores, there are 20480 cells (16 lat x 16 lon x 80 vertical levels) in each core. From results reported in the literature, the parallel efficiency of the coupled model is comparable to other ocean-alone or atmosphere-alone models when having similar number of grid points per CPU core [Marshall et al. 1997, Zhang et al. 2013].

P22 l14: Fluctuations in computational time might be resolved by performing a few runs and averaging the execution times. I am still convinced that idle ocean points when using many cores are seriously affecting load balancing and thus overall performance when using the system sequentially.

Reply: Yes, we agree with the reviewer that the fluctuation of the execution times may be due to the idle ocean points. However, Section 5 aims to show the coupled code has good scalability and the coupler does not slow down the simulation. Hence, we have revised our manuscript and removed the discussion on the fluctuation of CPU time.

P23 l9: 'We hypothesize' not 'we hypothesis'.

Reply: We have now fixed this typo.

P23 l11: Your results suggest that using more cores increases the cost of coupling. That is worrysome. Is there an upper limit to this cost?

Reply: Yes, we performed clean tests on another cluster (Shaheen-II in KAUST) and turned off unnecessary I/O. Our old COMPAS cluster is down and we cannot use it. We found that the cost of the coupling will increase only by 3% and the total time spent on the coupler does not increase. We have updated our manuscript.

P24 Table 3: For a clean evaluation of ESMF cost, switch off I/O, because it is handled by each component. Also, use the built-in ESMF Clock, which will tell you exactly where time is being spent.

Reply: We have performed a clean evaluation of ESMF cost by turning off unnecessary I/O. We have updated our results in Section 5 and we have shown that the coupled code does not slow down the ocean and atmosphere simulations.

SKRIPS v1.0: A regional coupled ocean–atmosphere modeling framework (MITgcm–WRF) using ESMF/NUOPC, description and preliminary results for the Red Sea

Rui Sun[1], Aneesh C. Subramanian[1], Arthur J. Miller[1], Matthew R. Mazloff[1], Ibrahim Hoteit[2], and Bruce D. Cornuelle[1]

[1]Scripps Institution of Oceanography, La Jolla, California, USA
[2]Physical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Correspondence: Rui Sun (rus043@ucsd.edu); Aneesh Subramanian (acsubram@ucsd.edu)

Abstract. A new regional coupled ocean–atmosphere model is developed ~~to study air–sea feedbacks~~and its implementation is presented in this paper. The coupled model is based on two open-source community model components: (1) MITgcm ocean model; (2) Weather Research and Forecasting (WRF) atmosphere model. The coupling between these components is performed using ESMF (Earth System Modeling Framework) and implemented according to National United Operational Prediction Capability (NUOPC) ~~consortium~~protocols. The coupled model is named the Scripps–KAUST Regional Integrated Prediction System (SKRIPS). The SKRIPS ~~allows affordable regional simulation of oceanic mixed layer heat and momentum interactions with atmospheric boundary layer dynamics at mesoscale and higher resolution. This can capture the feedbacks which are not well-resolved in coarse-resolution global coupled models and are absent in regional uncoupled models. After the model was created and passed a typical suite of consistency checks, we demonstrated it using a~~ is validated with a real-world example ~~. It simulated~~ by simulating a 30-day period including a series of ~~heat wave events that occurred~~ extreme heat events occurring on the eastern shore of the Red Sea region in June 2012. The results obtained by using the coupled model, along with those in forced ~~uncoupled ocean or atmosphere model~~ stand-alone oceanic or atmospheric simulations, are compared with observational ~~and reanalysis data. All~~ data and reanalysis products. We demonstrate that the coupled model is capable of performing coupled ocean–atmosphere simulation, although all configurations of coupled and uncoupled models have good skill in modeling ~~variables of interest in the region. The coupled model shows improved skill in temperature and circulation evaluation metrics.~~ the heat events. In addition, a scalability test is performed to investigate the parallelization of the coupled model. The results indicate that the ~~ESMF/NUOPC interface~~ coupled model code scales well and the ESMF/NUOPC coupler accounts for less than ~~10~~5% of the total computational resources ~~used in the simulation~~in the Red Sea test case. The coupled model, documentation, and tutorial cases used in this work are available at https://library.ucsd.edu/dc/collection/bb1847661c, and the source code is maintained at https://github.com/iurnus/scripps_kaust_model.

1 Introduction

Accurate and efficient forecasting of oceanic and atmospheric circulation is essential for a wide variety of high-impact societal needs, including extreme weather and climate events (Kharin and Zwiers, 2000; Chen et al., 2007), environmental protection and coastal management (Warner et al., 2010), management of fisheries (Roessig et al., 2004), marine conservation (Harley et al., 2006), water resources (Fowler and Ekström, 2009), and renewable energy (Barbariol et al., 2013). Effective forecasting relies on high model fidelity and accurate initialization of the models with the observed state of the coupled ocean–atmosphere system. Although global coupled models are now being implemented with increased resolution, higher-resolution regional coupled models, if properly driven by the boundary conditions, can provide an affordable way to study air–sea feedback for frontal-scale processes.

A number of regional coupled ocean–atmosphere models have been developed for various goals in the past decades. An early example of building a regional coupled model for realistic simulations focused on accurate weather forecasting in the Baltic Sea (Gustafsson et al., 1998; Hagedorn et al., 2000; Doscher et al., 2002), and ~~suggested~~ showed that the coupled model improved the SST (Sea Surface Temperature) and atmospheric circulation forecast. Enhanced numerical stability in the coupled simulation was also observed. These early attempts were followed by other practitioners in ocean-basin-scale climate simulations (e.g. Huang et al., 2004; Aldrian et al., 2005; Xie et al., 2007; Seo et al., 2007; Somot et al., 2008; Fang et al., 2010; Boé et al., 2011; Zou and Zhou, 2012; Gualdi et al., 2013; Van Pham et al., 2014; Chen and Curcic, 2016; Seo, 2017). For example, Huang et al. (2004) implemented a regional coupled model to study three major important patterns contributing to the variability and predictability of the Atlantic climate. The study suggested that these patterns originate from air–sea coupling within the Atlantic Ocean or by the oceanic responses to atmospheric internal forcing. Seo et al. (2007) studied the nature of ocean–atmosphere feedbacks in the presence of oceanic mesoscale eddy fields in the eastern Pacific Ocean sector. The evolving SST fronts were shown to drive an unambiguous response of the atmospheric boundary layer in the coupled model, and lead to model anomalies of wind stress curl, wind stress divergence, surface heat flux, and precipitation that resemble observations. This study helped substantiate the importance of ocean–atmosphere feedbacks involving oceanic mesoscale variability features.

In addition to basin-scale climate simulations, regional coupled models are also used to study weather extremes. For example, the COAMPS (Coupled Ocean/Atmosphere Mesoscale Prediction System) was applied to investigate idealized tropical cyclone events (Hodur, 1997). This work was then followed by other realistic extreme weather studies. ~~For example ,~~ Another example is the investigation of extreme bora wind events in the Adriatic Sea ~~were investigated~~ using different regional coupled models (Loglisci et al., 2004; Pullen et al., 2006; Ricchi et al., 2016). The coupled simulation results demonstrated improvements in describing the air–sea interaction processes by taking into account ocean surface heat fluxes and wind-driven ocean surface wave effects (Loglisci et al., 2004; Ricchi et al., 2016). It was also found in model simulations that SST after bora wind events had a stabilizing effect on the atmosphere, reducing the atmospheric boundary layer mixing and yielding stronger near-surface wind (Pullen et al., 2006). Regional coupled models were also used for ~~improving~~ studying the forecasts of ~~the hurricane pathand intensity, predicting~~ hurricanes, including hurricane path, hurricane intensity, SST variation, and ~~forecasting wind speeds~~wind speed (Bender and Ginis, 2000; Chen et al., 2007; Warner et al., 2010).

Regional coupled modeling systems also play important roles in studying the effect of surface variables (e.g., surface evaporation, precipitation, surface roughness) in the coupling processes of ~~ocean~~ oceans or lakes. One example is the study conducted by Powers and Stoelinga (2000), who developed a coupled model and investigated the atmospheric frontal passages over the Lake Erie region. Sensitivity analysis was performed to demonstrate that taking into account lake surface roughness parameterization in the atmosphere model can improve the calculation of wind stress and heat flux. Another example is the investigation by Turuncoglu et al. (2013), who compared a regional coupled model with uncoupled models and demonstrated the improvement of the coupled model in capturing the response of Caspian Sea levels to climate variability.

In the past ten years, many regional coupled models have been developed using modern model toolkits (Zou and Zhou, 2012; Turuncoglu et al., 2013; Turuncoglu, 2019) and include waves (Warner et al., 2010; Chen and Curcic, 2016), sediment transport (Warner et al., 2010), sea ice (Van Pham et al., 2014), and chemistry packages (He et al., 2015). However, it is still desirable and useful to develop a new coupled regional ocean–atmosphere model implemented using an efficient coupling framework and with state estimation capabilities. The goal of this work is to (1) introduce the design of a newly developed regional coupled ocean–atmosphere modeling system, (2) describe the implementation of the modern coupling framework, (3) ~~present preliminary simulation results in the Red Sea region~~validate the coupled model using a real-world example, and (4) demonstrate and discuss the parallelization of the coupled model. In the coupled system, the oceanic model component is the MIT general circulation model (MITgcm) (Marshall et al., 1997) and the atmospheric model component is the Weather Research and Forecasting (WRF) model (Skamarock et al., 2005). To couple the model components in the present work, the Earth System Modeling Framework (ESMF) (Hill et al., 2004) is used because of its advantages in conservative re-gridding capability, calendar management, logging and error handling, and parallel communications. The National United Operational Prediction Capability (NUOPC) layer in ESMF (Sitz et al., 2017) is also used ~~(Sitz et al., 2017). The additional NUOPC wrapper layer between coupled model and ESMF simplifies the implementations~~ between model components and ESMF. Using the NUOPC layer can simplify the implementation of component synchronization, execution, and other common tasks in the coupling. The innovations in our work are: (1) we use ESMF/NUOPC, which is a community supported computationally efficient coupling software for earth system models, and (2) we ~~used~~ use MITgcm together with WRF. The resulting coupled model is being developed as a coupled forecasting tool for coupled data assimilation and subseasonal to seasonal (S2S) forecasting. By coupling WRF and MITgcm for the first time with ESMF, we can provide an alternative regional coupled model resource to a wider community of users. These atmospheric and oceanic model components have an active and well-supported user-base.

After ~~testing of~~ implementing the new coupled model, we demonstrate it on a series of heat ~~wave~~ events that occurred on the eastern shore of the Red Sea region in June 2012. The simulated surface variables of the Red Sea (e.g., sea surface temperature, 2-m temperature, and surface heat fluxes) are examined and validated against available observational data and reanalysis products. To ~~assess the improvements gained from the coupled simulation~~demonstrate the coupled model can perform coupled ocean–atmosphere simulations, the results are compared with those obtained using stand-alone oceanic or atmospheric models. This paper focuses on the technical aspects of the SKRIPS, and is not a full investigation of the importance of coupling for these extreme events~~, which is outside of the scope of this paper, which focuses on the technical aspects~~. In addition, a scalability test of the coupled model is performed to investigate its parallel capability.

The rest of this paper is organized as follows. The description of the individual modeling components and the design of the coupled modeling system are detailed in Section 2. Section 3 introduces the ~~experimental design , validation data, and analysis methodology~~design of validation experiment and the validation data. Section 4 discusses the ~~results obtained from the coupled model~~preliminary results in the validation test. Section 5 details the parallelization test of the coupled model. The last section concludes the paper and presents an outlook for future work.

2 Model Description

The newly developed regional coupled modeling system is introduced in this section. The general design of the coupled model, descriptions of individual components, and ESMF/NUOPC coupling framework are presented below.

2.1 General design

The schematic description of the coupled model is shown in Fig. 1~~(a)~~. The coupled model is comprised of five components: oceanic component MITgcm, atmospheric component WRF, MITgcm–ESMF interface, WRF–ESMF interface, and ESMF ~~/NUOPC~~ coupler. They are to be detailed in the following sections.

The coupler component runs in both directions: (1) from WRF to MITgcm, and (2) from MITgcm to WRF. From WRF to MITgcm, the coupler collects the surface atmospheric variables (i.e., ~~solar radiation~~radiative flux, turbulent heat flux, wind velocity, precipitation, evaporation) from WRF and updates the surface forcing ~~variables (net~~ (i.e., net surface heat flux, wind stress, freshwater flux) to drive MITgcm. From MITgcm to WRF, the coupler collects ocean surface variables (i.e., SST and ocean surface velocity~~from the MITgcm and uses them as the surface boundary conditionin WRF~~) from MITgcm and updates them in WRF as the bottom boundary condition. Re-gridding the data from either model component ~~will be~~ is performed by the coupler, in which various coupling intervals and schemes can be specified by ~~the~~ ESMF (Hill et al., 2004).

~~The general code structure and run sequence of the coupled ocean–atmosphere model. In panel (a), the black block is the *application driver*; the red block is the *parent gridded component* called by the *application driver*; the green/blue blocks are the *child gridded/coupler components* called by the *parent gridded component*. In panel (b), OCN, ATM, and CON denote oceanic component, atmospheric component and connector component, respectively. The red arrows indicate the model components are sending data to the connector and the yellow arrows indicate the model components are reading data from the connector. The horizontal black arrows indicate the time axis of each component and the ticks on the time axis indicate the coupling time step.~~

2.2 MITgcm Ocean Model

The MITgcm (Marshall et al., 1997) is a 3-D, finite-volume, general circulation model used by a broad community of researchers for a wide range of applications at various spatial and temporal scales. The model code and documentation, which are under continuous development, are available on the MITgcm webpage http://mitgcm.org/. The 'Checkpoint 66h' (June 2017) version of MITgcm is used in the present work.

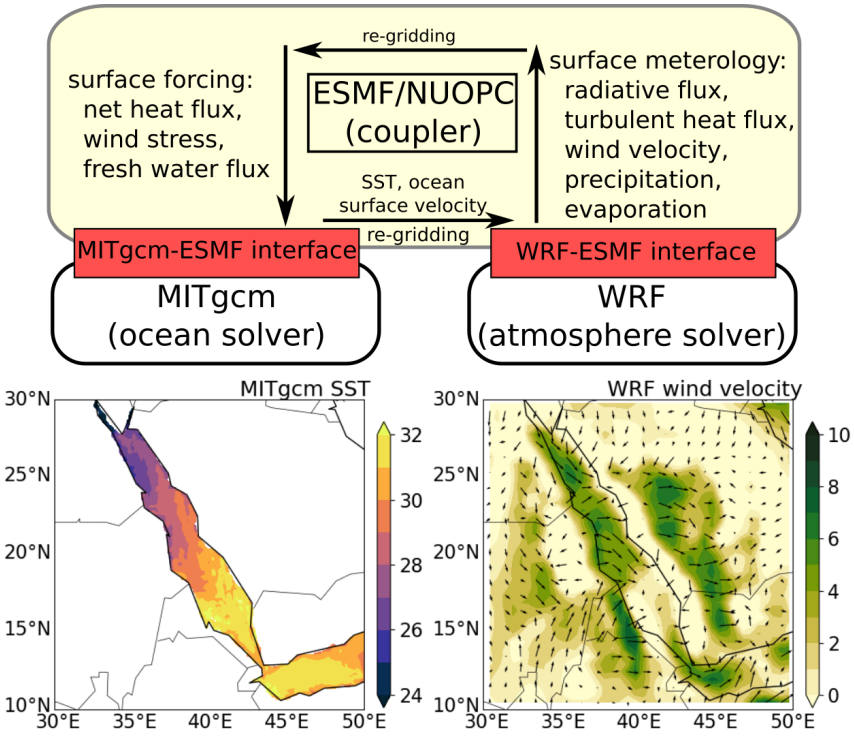


Figure 1. The schematic description of the coupled ocean–atmosphere model. The white blocks are the oceanic and atmospheric components; the red blocks are the implemented MITgcm–ESMF and WRF–ESMF interfaces; the yellow block is the ESMF/NUOPC coupler. From WRF to MITgcm, the coupler collects the surface atmospheric variables (i.e., radiative flux, turbulent heat flux, wind velocity, precipitation, evaporation) and updates the surface forcing (i.e., net surface heat flux, wind stress, freshwater flux) to drive MITgcm. From MITgcm to WRF, the coupler collects ocean surface variables (i.e., SST and ocean surface velocity) and updates them in WRF as the bottom boundary condition.

The MITgcm is designed to run on high-performance computing (HPC) platforms and can run in non-hydrostatic and hydrostatic modes. It integrates the primitive (Navier-Stokes) equations, under the Boussinesq approximation, using finite volume method on a staggered 'Arakawa C-grid'. The MITgcm uses modern physical parameterization schemes for subgrid-scale horizontal and vertical mixing and tracer properties. The code configuration includes build-time C pre-processor (CPP) options and run-time switches, which allow for great computational modularity in MITgcm to study a variety of oceanic phenomena (Evangelinos and Hill, 2007).

To implement the MITgcm–ESMF interface, we ~~separated~~ separate the MITgcm main program into three subroutines that handle initialization, running, and finalization, shown in Fig. 2(a). These subroutines are used by the ESMF/NUOPC coupler that controls the oceanic component in the coupled run. The surface boundary fields on the ocean surface ~~is~~ are ex-

changed online[1] via the MITgcm–ESMF interface during the simulation. The MITgcm ~~SST and ocean surface velocity~~ ocean surface variables are the export boundary fields, and the atmospheric surface ~~forcing~~ variables are the import boundary fields (see Fig. 2(b)). These boundary fields are registered in the coupler following NUOPC ~~consortium and timestamps~~protocols and timestamps[2] are added to them for the coupling. In addition, MITgcm grid information is also provided to the coupler in the initialization subroutine for online re-gridding of the exchanged boundary fields. To carry out the ~~high-resolution simulation~~ coupled simulation on HPC clusters, the MITgcm–ESMF interface runs in parallel via MPI communications. The ~~implementations~~ implementation of the present MITgcm–ESMF interface ~~are~~ is based on the baseline MITgcm–ESMF ~~coupler~~interface (Hill, 2005), but ~~we updated it to couple~~ updated for compatibility with the modern version of ESMF/NUOPC~~with MITgcm~~. We also ~~modified the baseline coupler~~ modify the baseline interface to receive atmosphere surface fluxes and send ocean surface variables~~(i. e., SST and ocean surface velocity).~~ .

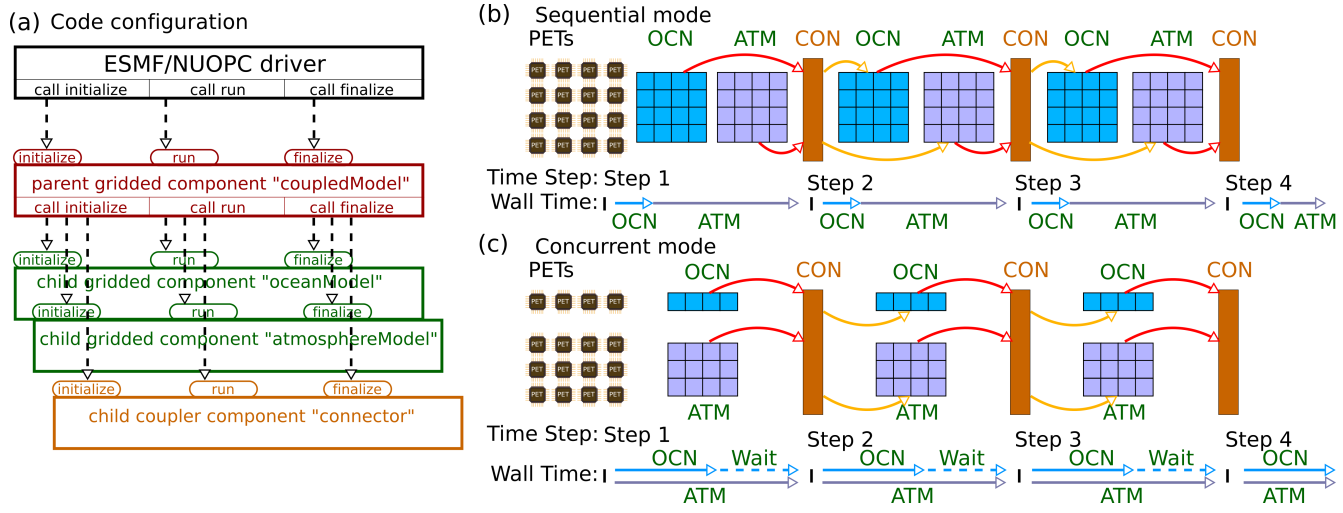


Figure 2. The general code structure and run sequence of the coupled ocean–atmosphere model. In panel (a), the black block is the *application driver*; the red block is the *parent gridded component* called by the *application driver*; the green/brown blocks are the *child gridded/coupler components* called by the *parent gridded component*. Panel (b) and (c) shows the sequential and concurrent mode implemented in SKRIPS, respectively. PETs (Persistent Execution Threads) are single processing units (e.g., CPU or GPU cores) defined by ESMF. OCN, ATM, and CON denote oceanic component, atmospheric component and connector component, respectively. The blocks under PETs are the CPU cores in the simulation; the small blocks under OCN or ATM are the small sub-domains in each core; the block under CON is the coupler. The red arrows indicate the model components are sending data to the connector and the yellow arrows indicate the model components are reading data from the connector. The horizontal arrows indicate the time axis of each component and the ticks on the time axis indicate the coupling time step.

[1]In this manuscript, 'online' means the manipulations are performed via subroutine calls during the execution of the simulations; 'offline' means the manipulations are performed when the simulations are not executing.

[2]In ESMF, 'timestamp' is a sequence of ~~number~~numbers, usually based on the time, to identify ~~the~~ ESMF fields. Only the ESMF fields having the correct timestamp will be transferred in the coupling.

2.3 WRF Atmospheric Model

The Weather Research and Forecasting (WRF) Model (Skamarock et al., 2005) is developed by NCAR/MMM (Mesoscale and Microscale Meteorology Division). It is a 3-D, finite-difference atmospheric model with a variety of physical parameterizations of sub-grid scale processes for predicting a broad spectrum of applications. WRF is used extensively for operational forecasts (http://www.wrf-model.org/plots/wrfrealtime.php) as well as realistic and idealized dynamical studies. The WRF code and documentation are under continuous development on Github (https://github.com/wrf-model/WRF).

In the present work, the Advanced Research WRF dynamic version (WRF-ARW, version 3.9.1.1, https://github.com/NCAR/WRFV3/releases/tag/V3.9.1.1) is used. It solves the compressible Euler non-hydrostatic equations, and also includes a run-time hydrostatic option. The WRF-ARW uses a terrain-following hydrostatic pressure coordinate system in the vertical direction and utilizes the 'Arakawa C-grid'. WRF incorporates various physical processes including microphysics, cumulus parameterization, planetary boundary layer, surface layer, land surface, and longwave~~and~~ /shortwave radiations, with several options available for each process.

Similar ~~with the implementations~~ to the implementation in MITgcm, WRF is also separated into initialization, run, and finalization subroutines to enable the WRF–ESMF interface to control the atmosphere model during the coupled simulation, shown in Fig. 2(a). The implementation of the present WRF–ESMF interface is based on the prototype interface (Henderson and Michalakes, 2005). In the present work, the prototype WRF–ESMF interface is updated to ~~a modern version~~ modern versions of WRF-ARW and ~~a modern version of~~ ESMF, based on the NUOPC layer. This prototype interface is also expanded to interact with the ESMF/NUOPC coupler to receive the ocean surface variables and send the atmosphere surface fluxes. The surface boundary condition fields are registered in the coupler following the NUOPC ~~consortium~~ protocols with timestamps. The WRF grid information is also provided for online re-gridding by ESMF. To carry out the ~~high-resolution simulation~~ coupled simulation on HPC clusters, the WRF–ESMF interface also runs in parallel via MPI communications.

2.4 ESMF/NUOPC Coupler

The coupler is implemented using ESMF version 7.0.0. The ESMF is selected because of its high-performance and flexibility for building and coupling weather, climate, and related Earth science applications (Collins et al., 2005; Turuncoglu et al., 2013; Chen and Curcic, 2016; Turuncoglu and Sannino, 2017). It has a superstructure for representing the model and coupler components and an infrastructure of commonly used utilities, including conservative grid remapping, time management, error handling, and data communications.

The general code structure of the coupler is shown in Fig. 2. To build the ESMF/NUOPC driver, a main program is implemented to control an ESMF parent component, which controls the child components. In the present work, three child components are implemented: (1) the oceanic component; (2) the atmospheric component; and (3) the ESMF coupler. The coupler is used here because it performs the two-way interpolation and data transfer (Hill et al., 2004). In ESMF, the model components can be run in parallel as a group of Persistent Execution Threads (PETs), which are single processing units (~~i.e.CPU, GPU~~e.g.,

CPU or GPU cores) defined by ESMF. In the present work, the PETs are created according to the grid decomposition, and each PET is associated with an MPI process~~running on a separate processor~~.

The ESMF ~~also~~ allows the PETs ~~running~~ to run in sequential mode, concurrent mode, or ~~a mixed mode . We selected the sequential mode in the implementations~~mixed mode (for more than three components). We implemented both sequential and concurrent modes in SKRIPS, shown in Fig. 2(b) and 2(c). In sequential mode, a set of ESMF gridded/coupler components ~~runs~~ are run in sequence on the same set of PETs. At each coupling time step, the oceanic component is executed when the ~~atmosphere~~ atmospheric component is completed or vice versa. ~~However~~On the other hand, in concurrent mode, the gridded components are created and run on mutually exclusive sets of PETs. ~~There are some advantages of concurrent mode, however the simplicity of sequential mode makes it a natural starting point (Collins et al., 2005), and it is chosen for this work.~~

If one component finishes earlier than the other, its PETs are idle and have to wait for the other component, shown in Fig. 2(c). However the PETs can be optimally distributed by the users to best achieve load balance.

In ESMF, the gridded components are used to represent models, and coupler components are used to connect these models. The interfaces and data structures in ESMF have few constraints, providing the flexibility to be adapted to many modeling systems. However, the flexibility of the gridded components can limit the interoperability across different modeling systems. To address this issue, the NUOPC layer is developed to provide the coupling conventions and the generic representation of the model components (e.g. drivers, models, connectors, mediators). The NUOPC layer in the present coupled model is implemented according to ~~the~~ consortium documentations (Hill et al., 2004; Theurich et al., 2016), and the oceanic/atmospheric component each has:

1. Prescribed variables for NUOPC to link the components;
2. The entry point for registration of the components;
3. An *InitializePhaseMap* which describes a sequence of standard initialization phases, including ~~advertising~~ documenting the fields that a component can provide, checking and mapping the fields to each other, and initializing the fields that will be used;
4. A *RunPhaseMap* that checks the incoming clock of the driver, examines the timestamps of incoming fields, and runs the component;
5. Timestamps on exported fields consistent with the internal clock of the component;
6. The *finalization method* to clean up all allocations.

The subroutines that handle initialization, running, and finalization in MITgcm and WRF ~~will be~~ are included in the *InitializePhaseMap*, *RunPhaseMap*, and *finalization method* in the NUOPC layer, respectively.

3 Experiment Design and Observational Datasets

~~To test the coupled model, we applied it to study~~ We simulate a series of heat ~~wave~~ events in the Red Sea region~~. We selected the extreme heat wave events because of their societally relevant impacts.~~ , with a focus on validating and assessing the technical aspects of the coupled model. There is a desire for improved and extended forecasts in this region, and future work will investigate whether a coupled framework can advance this goal. The extreme heat events are chosen as a test case due to their societal importance. While these events and the analysis here do not highlight the value of coupled forecasting, these real-world events are adequate to demonstrate the performance and physical realism of the coupled model code implementation. The simulation of the Red Sea extends from 0000 UTC 01 June 2012 to 0000 UTC 01 July 2012. We select this month because of the record-high surface air temperature observed in the Makkah region, located 70 km inland from the eastern shore of the Red Sea (Abdou, 2014).

The computational domain and bathymetry are shown in Fig. 3. The model domain is centered at 20° N and 40° E, and the bathymetry is from the 2-minute Gridded Global Relief Data (ETOPO2) (National Geophysical Data Center, 2006). WRF is implemented using a horizontal grid of 256×256 points and grid spacing of 0.08°~~, using~~ . The cylindrical equidistant map (latitude-longitude) projection is used. There are 40 terrain-following vertical levels, more closely spaced in the atmospheric boundary layer. The time step for atmosphere simulation is 30 seconds, which is to avoid violation of the CFL condition. The Morrison 2-moment scheme (Morrison et al., 2009) is used to resolve the microphysics. The updated version of the Kain–Fritsch convection scheme (Kain, 2004) is used with the modifications to include the updraft formulation, downdraft formulation, and closure assumption. The Yonsei University (YSU) scheme (Hong et al., 2006) is used for the planetary boundary layer (PBL), and the Rapid Radiation Transfer Model for GCMs (RRTMG; Iacono et al. (2008)) is used for longwave and shortwave radiation transfer through the atmosphere. The Rapid Update Cycle (RUC) land surface model is used for the land surface processes (Benjamin et al., 2004). The MITgcm uses the same horizontal grid spacing as WRF, with 40 vertical z-levels that are more closely spaced near the surface. The time step of the ocean model is 120 seconds. The horizontal sub-grid mixing is parameterized using nonlinear Smagorinsky viscosities, and the K-profile parameterization (KPP) (Large et al., 1994) is used for vertical mixing processes.

In the coupling process, the ocean model sends SST and ocean surface velocity to the coupler, and they are used directly as the boundary conditions in the atmosphere model. The atmosphere model sends the surface fields to the coupler, including (1) net surface shortwave/longwave radiation, (2) surface latent/sensible heat flux, (3) 10-m wind speed, (4) ~~net~~ precipitation, (5) evaporation. The ocean model uses the atmosphere surface fields to compute the surface forcing, including (1) total net surface heat flux, (2) surface wind stress, (3) freshwater flux. The total net surface heat flux is computed by adding latent heat flux, sensible heat flux, and net surface shortwave/longwave radiation fluxes. The surface wind stress is computed by using the 10-m wind speed (Large and Yeager, 2004). The freshwater flux is the difference between precipitation and evaporation. The latent and sensible heat fluxes are computed by using COARE 3.0 bulk algorithm in WRF (Fairall et al., 2003). In the coupled code, different bulk formulae in WRF or MITgcm can also be used.

To ~~study the air–sea interactions~~validate the coupled model, the following sets of simulations using different surface forcings are performed according to the tests (Warner et al., 2010; Turuncoglu et al., 2013; Ricchi et al., 2016):

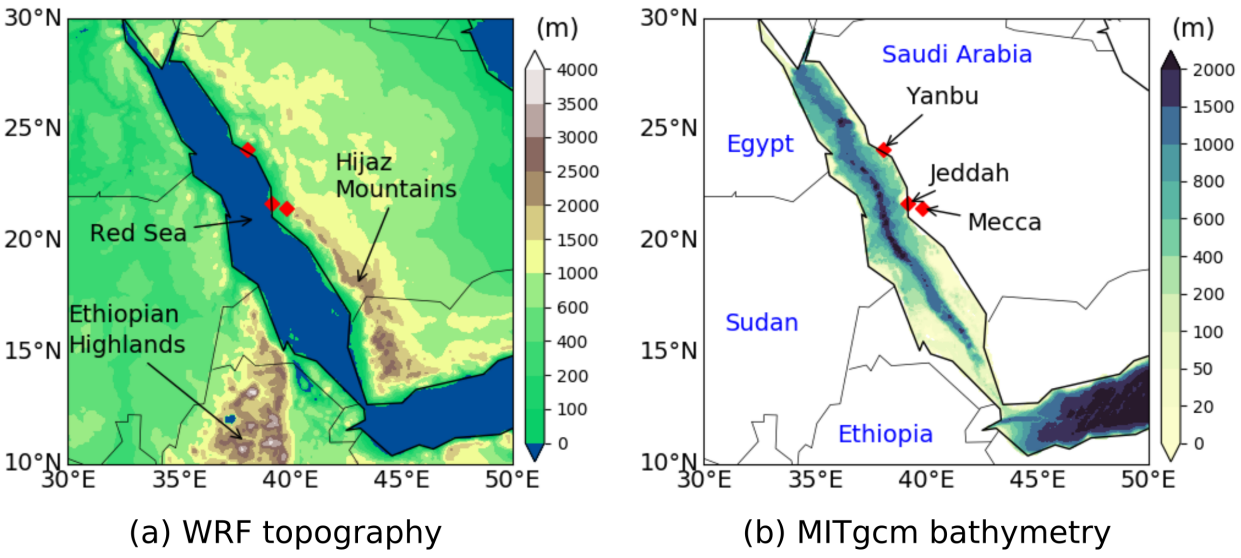


Figure 3. The WRF topography and MITgcm bathymetry in the simulations. Three major cities near the eastern shore of the Red Sea are highlighted. The Hijaz Mountains and Ethiopian Highlands are also highlighted.

1. Run CPL: a two-way coupled MITgcm–WRF simulation. The coupling interval is 20 minutes to capture the diurnal cycle (Seo et al., 2014). This run tests the ~~performance~~ implementation of the two-way coupled ocean–atmosphere model.

2. Run ATM.STA: a stand-alone WRF simulation with its initial SST kept constant throughout the simulation. This run allows ~~assesment~~ assessment of the WRF model behavior with realistic, but persistent SST. This case serves as a benchmark to highlight the difference between coupled and uncoupled runs, and also to demonstrate the impact of evolving SST.

3. Run ATM.DYN: a stand-alone WRF simulation with a varying, prescribed SST based on HYCOM/NCODA reanalysis data. This allows assessing the WRF model behavior with updated ~~sea surface temperature. The ocean's effect on the atmosphere is considered in the ATM. DYN run. In~~ SST and is used to validate the coupled model. It is noted that in practice an accurately evolving SST would not be available for forecasting~~, however the comparison between ATM. DYN and CPL runs is used to demonstrate skill in the coupled model.~~ .

4. Run OCN.DYN: a stand-alone MITgcm simulation forced by the ERA5 ~~dataset~~reanalysis data. The bulk formula in MITgcm is used to derive the turbulent heat fluxes. This run assesses the MITgcm model behavior with prescribed lower-resolution atmospheric surface forcing, and ~~like the ATM.DYN run is used to show the skill of the~~ is also used to validate the coupled model.

The ocean model uses the ~~assimilated~~ data assimilating HYCOM/NCODA $1/12^{\circ}$ global ~~analysis~~ reanalysis data as initial and boundary conditions for ocean temperature, salinity, and horizontal velocities (https://www.hycom.org/dataserver/gofs-3pt1/reanalysis). The boundary conditions for the ocean are updated on a ~~daily~~ 3-hourly basis and linearly interpolated between two simulation time steps. A sponge layer is applied at the lateral boundaries, with a thickness of 3 grid cells~~and inner /~~. The inner and outer boundary relaxation timescales of the sponge layer are 10 ~~/~~and 0.5 days, respectively. In CPL, ATM.STA, and ATM.DYN runs, we ~~used~~ use the same initial condition and lateral boundary condition for the atmosphere. The atmosphere is initialized using the ECMWF ERA5 reanalysis ~~dataset~~data, which has a grid resolution of approximately 30 km (Hersbach, 2016). The same data also provide the boundary conditions for air temperature, wind speed, and air humidity every ~~6~~ 3 hours. The atmosphere boundary conditions are also linearly interpolated between two simulation time steps. The lateral boundary values are specified in WRF in the 'specified' zone, and the 'relaxation' zone is used to nudge the solution from the domain toward the boundary condition value. Here we used the default width of one point for the specific zone and four points for the relaxation zone. The pressure at the top of the atmosphere is 50 hPa. In ATM.STA run, the SST from ~~the~~ HYCOM/NCODA ~~data~~ is used as initial and persistent SST. The time-varying SST in ATM.DYN run is also generated using HYCOM/NCODA data. We ~~selected~~ select HYCOM/NCODA data because the ocean model initial condition and boundary conditions are generated using it. For the OCN.DYN run we select ~~the~~ ERA5 ~~dataset~~ data for prescribed atmospheric state because it also provides the atmospheric initial and boundary conditions in the CPL run. The initial condition, boundary condition, and forcing terms of ~~this run~~ all simulations are summarized in Table 1.

Table 1. The initial condition, boundary condition and forcing terms used in present simulations.

	initial and boundary conditions	ocean surface conditions	atmospheric forcings
CPL	ERA5 (atmosphere) HYCOM/NCODA (ocean)	from MITgcm	from WRF
ATM.STA	ERA5	HYCOM/NCODA initial condition kept constant	N.A.
ATM.DYN	ERA5	HYCOM/NCODA updated every ~~24~~ 3 hours	N.A.
OCN.DYN	HYCOM/NCODA	N.A.	ERA5 ~~+ MITgcm bulk formula~~

The ~~analysis of the results~~ validation of the coupled model focuses on temperature, heat flux, ~~surface wind, and evaporation~~and surface wind. Our aim is to validate the coupled model and show that the heat and momentum fluxes simulated by the coupled model are comparable to the observations or the reanalysis data. The simulated SST data are validated against the OSTIA (Operational Sea Surface Temperature and Sea Ice Analysis) system in GHRSST (Group for High Resolution Sea Surface Temperature) (Donlon et al., 2012; Martin et al., 2012)~~, and the~~ . The simulated SST is also validated against HYCOM/NCODA data to show the increase of the error. The simulated 2-meter air temperature (T2) ~~is validated against the ECMWF~~ fields

are validated using ERA5~~dataset. To evaluate the modeling of the heat wave event~~ . In addition, the T2 in three major cities near the eastern shore of ~~Red Sea, the diurnal temperature variation is compared with observed daily maximum and minimum temperatures~~ the Red Sea are validated using ERA5 and the ground observations from NOAA National Climate Data Center (NCDC climate data online at http://cdo.ncdc.noaa.gov/CDO/georegion). For this comparison the T2 fields from both the simulations and ERA5 are interpolated to the NCDC stations. When interpolating ERA5 data to the NCDC stations near the coast, only the data saved on ERA5 land points (land-sea mask>90%) are used in the bi-linear interpolation. The high/low temperature every 24 hours from the simulations and ERA5 are compared to the daily maximum/minimum temperatures with NCDC data. Surface heat fluxes (e.g., latent heat, sensible heat, longwave and shortwave radiations), which ~~are important for ocean–atmosphere interactions, are compared with~~ drives the oceanic component in the coupled model, are validated using MERRA-2 (Modern-Era Retrospective analysis for Research and Applications, version 2) ~~datasets~~data (Gelaro et al., 2017). The MERRA-2 ~~dataset is selected because~~ data are selected because (1) it is an independent reanalysis data compared to the initial and boundary conditions used in the simulations~~. The MERRA-2 dataset~~ , and (2) it also provides a $0.625^o \times 0.5^o$ (lon $\times$ lat) resolution reanalysis fields of turbulent heat fluxes~~.~~ (THF). The 10-m wind speed is also compared with MERRA-2 data to validate the momentum flux in the coupled code. To compare with validation data, we interpolated the validation data on the lower resolution grid to the higher resolution grid of the regional model. The validation data are summarized in Table 2. The validation of the freshwater flux is shown in the Appendix because (1) the evaporation is proportional to the latent heat in the model and (2) the precipitation is zero in three major cities near the coast in Fig. 3.

Table 2. The ~~dataset~~ observational data and reanalysis data used to validate the simulation results.

variable	validation data
sea surface temperature (SST)	GHRSST and HYCOM/NCODA
2-meter air temperature (T2)	ERA5 and NCDC climate data
turbulent heat fluxes	MERRA-2
~~solar radiations~~ radiative fluxes	MERRA-2
~~surface~~ 10-meter wind	MERRA-2
~~surface evaporation MERRA-2~~	

4 Results and Discussions

The Red Sea is an elongated basin covering the area between 12-30°N and 32-43°E. The basin is 2250 km long, extending from the Suez and Aqaba gulfs in the north to the strait of Bal el-Mandeb in the south, which connects the Red Sea and the Indian Ocean. ~~Since the global models with coarse resolution cannot properly resolve local features in the narrow basin of the Red Sea (Yao et al., 2014b, a; Zhan et al., 2014), regional models with relatively higher resolutions can be used as dynamical downscaling tools for extreme temperature studies (Li et al., 2018).~~ In this section, ~~results of the simulations~~ the simulation

results obtained by using different model configurations ~~will be presented and examined to assess the performance of the coupled model in simulating the heat wave events in the Red Sea region.~~ are presented to show that SKRIPS is capable of performing coupled ocean–atmosphere simulations.

4.1 2-meter Air Temperature (T2)

We begin our analysis by examining the simulated T2 from ~~various experiments. The simulation results obtained from coupled (CPL) run, the~~ the model experiments. Since the record-high temperature is observed in the Makkah region on June 2nd, the simulation results on June 2nd (36 or 48 hours after the initialization) are shown in Fig. 4. The ERA5 data, and ~~their associated difference are~~ the difference between CPL run and ERA5 are also 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~~ T2 patterns in the Red Sea region on June 2nd, compared with ~~the~~ ERA5 ~~dataset~~ in Fig. 4(II). Since ERA5 air temperature data are in good agreement with the NCDC ground observation data in the Red Sea region (~~comparison~~ detailed comparison of all stations are not shown), we use ERA5 data to validate the simulation results. The difference between the CPL run and ERA5 ~~dataset~~ is shown in Fig. 4(III). The ATM.STA and ATM.DYN simulation results ~~have consistent patterns with~~ are close to the CPL run results and thus are not shown, but their differences with respect to ~~the~~ ERA5 ~~data~~ are shown in Fig. 4(IV) and 4(V), respectively. Fig. 4(VI) to 4(X) show the ~~same~~ nighttime results after 48 hours. It can be seen in Fig. 4 that all simulations reproduce the T2 patterns over the Red Sea region reasonably well compared with ~~the~~ ERA5~~data~~. The mean T2 ~~differences~~ biases and RMSEs over the sea are ~~-1.55 °C (CPL), -1.66 °C (ATM. STA), and -1.70 °C (ATM.DYN) after 36 hours, and -0.99 °C (CPL), -1.10 °C (ATM.STA), and -1.12 °C (ATM.DYN) after 48 hours. The mean~~ shown in Table 3. The biases of the T2 ~~differences in all simulations are mostly the same compared with the mean and standard deviation of T2 (31.01 °C and 1.93 °C after 36 hours; 30.25 °C and 1.36 °C after 48 hours)~~are comparable with the biases reported in other WRF simulations for heat events (Imran et al., 2018). Fig. 4 also shows ~~that all simulations can capture the T2 diurnal variation~~ the diurnal variation of T2 in the Red Sea region, and ~~this~~ the diurnal variation will be further discussed later in this section.

The simulation results ~~for the heat wave events~~ on June 10th and 24th are shown in Fig. 5 to ~~demonstrate the performance of the~~ validate the coupled model over longer periods of time. In Fig. 5, we aim to show the difference over the sea to validate the coupled ocean–atmosphere model. It can be seen in Fig. 5(III) and 5(VIII) that the T2 patterns simulated by the coupled run are ~~consistent with the~~ generally consistent with ERA5~~dataset~~. The differences between ATM.STA and ATM.DYN simulation results with ~~respect to the~~ ERA5 ~~data~~ are shown in Fig. 5(IV), 5(V), 5(IX), and 5(X), respectively. It can be seen in the figure that the T2 errors on land are consistent for all three simulations. However, the T2 over the sea in CPL simulation has ~~a much smaller difference with the validation ERA5 data (10th: -1.02 °C; 24th: -0.84 °C)~~ smaller mean biases and RMSEs compared with the ATM.STA run~~(10th: -1.56 °C; 24th: -2.13 °C)~~, shown in Table 3. Although the difference of the biases is still very small compared with the mean T2 (~~31.12~~31.92 °C~~on 10th; 32.09 °C on 24th~~), the improvement of the coupled run ~~is comparible~~ on the 24th (1.02 °C) is comparable to the standard deviation of T2 (~~2.14~~1.64 °C~~on 10th; 2.02 °C on 24th~~). The ~~CPL run results are closer to the~~ T2 over the water in the CPL run is closer to ERA5 ~~dataset because the oceanic component (MITgcm) is providing updated~~ because MITgcm in the coupled model provides updated warming SST, which warms the T2; the ATM.STA

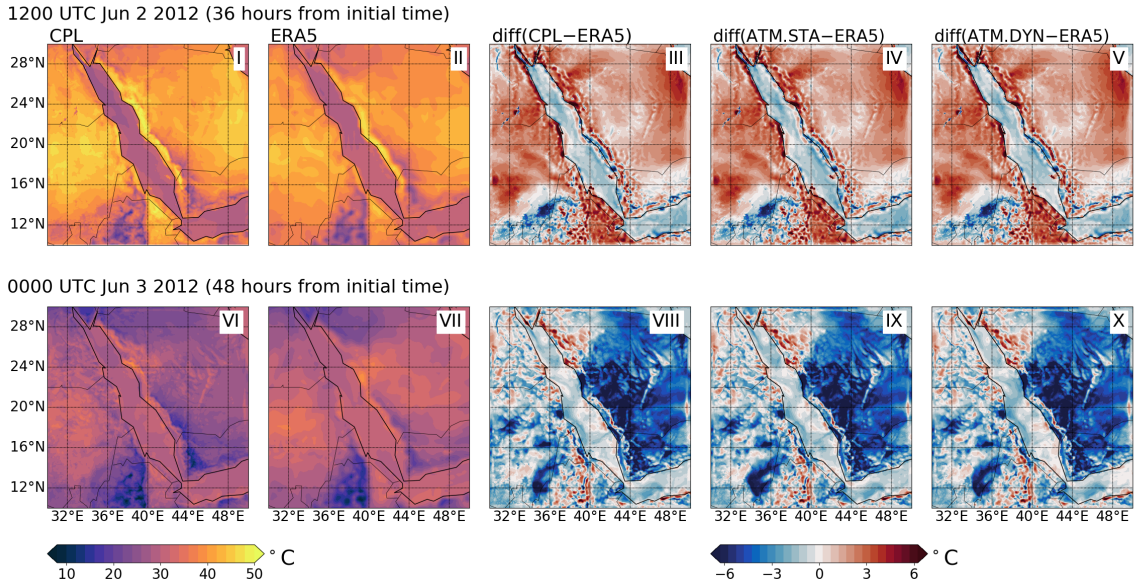


Figure 4. The ~~surface~~ 2-m air temperature as obtained from the CPL run, the ERA5 data, and their difference (CPL−ERA5). The ~~difference~~ differences between ATM.STA and ATM.DYN with ~~the~~ ERA5 ~~data~~ (i.e., ATM.STA−ERA5, ATM.DYN−ERA5) are also presented. The simulation initial time is 0000 UTC Jun 01 2012 for both snapshots. Two snapshots are selected: (1) 1200 UTC Jun 02 2012 (36 hours from initial time); (2) 0000 UTC Jun 03 2012 (48 hours from initial time). The results on Jun 02 are presented because the record-high temperature is observed in the Makkah region.

run uses a constant cooler SST from June 1st, and thus the T2 is determined by the constant cooler SST. On the other hand, when comparing the CPL run with the ATM.DYN run~~on June 24th, the difference is very small (-0.10~~, the mean difference is smaller (10th: +0.04 °C~~on June~~ ; 24th: -0.62 °C). This ~~is because the SST fields from CPL and~~ shows the CPL run is comparable to the ATM.DYN ~~runs are similar, which means that the SST in CPL run is tending to be similar to the realistic~~run which is driven by an updated warming SST.

~~To investigate~~

Table 3. The biases and RMSEs of the T2 simulated in all simulations in comparison with ERA5 data.

	after 36 hours	after 48 hours	after 9.5 days	after 23.5 days
CPL run	bias: -1.36; RMSE: 1.20	bias: -0.82; RMSE: 1.18	bias: -1.24; RMSE: 1.74	bias: -0.81; RMSE: 1.59
ATM.STA run	bias: -1.48; RMSE: 1.23	bias: -0.92; RMSE: 1.21	bias: -1.56; RMSE: 1.91	bias: -1.83; RMSE: 1.83
ATM.DYN run	bias: -1.36; RMSE: 1.21	bias: -0.84; RMSE: 1.18	bias: -1.20; RMSE: 1.46	bias: -1.43; RMSE: 1.37

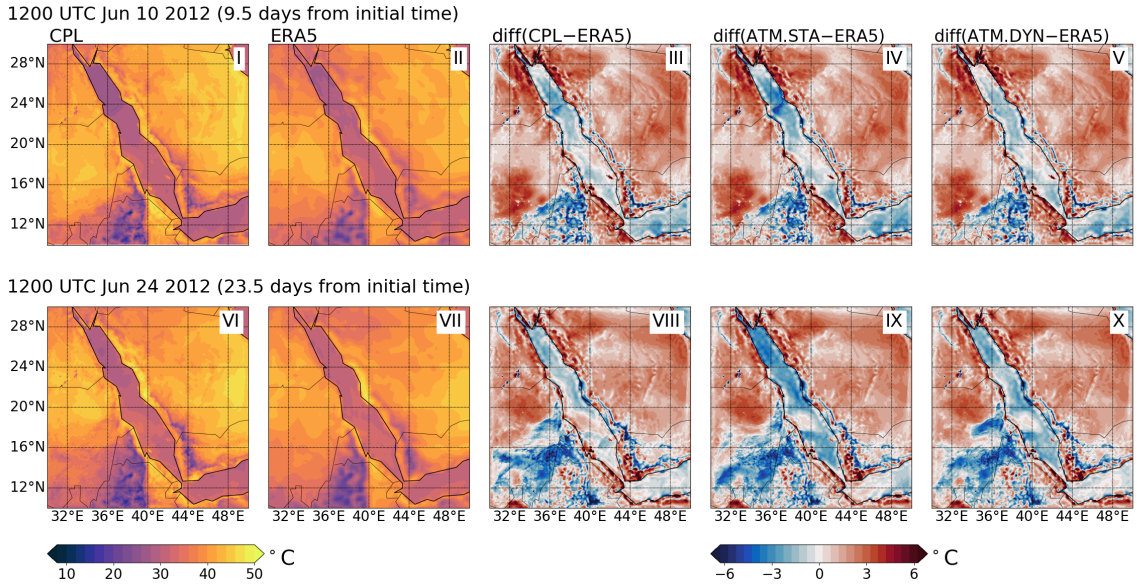


Figure 5. The ~~surface~~ 2-m air temperature as obtained from the CPL run, the ERA5 data, and their difference (CPL−ERA5). The difference between ATM.STA and ATM.DYN with ~~the~~ ERA5 data (i.e., ATM.STA−ERA5, ATM.DYN−ERA5) are also presented. The simulation initial time is 0000 UTC Jun 01 2012 for both snapshots. Two snapshots are selected: (1) 1200 UTC Jun 10 2012 (9.5 days from initial time); (2) 1200 UTC Jun 24 2012 (23.5 days from initial time).

To validate the diurnal T2 variation of the coupled model in Fig. 4, the time series of T2 in three major cities as simulated in CPL and ATM.STA runs are plotted in Fig. 6, starting from June 1st~~; the mean and standard deviation are shown in Fig. ??.~~. The ATM.DYN run results are similar ~~with~~ to the CPL run results and thus are not shown. To validate the simulation results, the time series in ERA5 data and the daily observed high/low temperature data from NOAA ~~National Climate Data Center~~ NCDC are also plotted. It can be seen ~~that~~ both coupled and uncoupled simulations generally captured the four major heat ~~waves~~ events (i.e., June 2nd, 10th, 17th, and 24th) and the T2 variations during the 30-day simulation~~are all captured by the simulations. Before June 17th (lead time < 16 days), the CPL and ATM.STA runs results are in good agreement with the ground observation and ERA5 dataset. The~~. For the daily high T2, the root mean square error (RMSE) ~~between the simulations and ground observation are 2.79~~in the CPL run (2.09 °C~~and 2.83 °C for CPL and~~) is close to the ATM.STA ~~runs, respectively. However, the error after June 18th (simulation lead time > 17 days)is larger for both CPL (3.42~~run (2.16 °C), and the error does not increase in the 30-day simulation. For the daily low T2, before June 20th (lead time < 19 days), the CPL and ATM.STA runs have consistent RMSEs compared with ground observation (CPL: 4.23 °C~~) and~~ ; ATM.STA~~(3.94~~: 4.39 °C)~~runs. It can be also seen that~~ . In Jeddah and Yanbu, the CPL run ~~better captures the daily high temperatures in Yanbu (RMSE difference: 2.77~~has better captured the daily low T2 after June 20th in CPL run (Jeddah: 3.95 °C~~) than ERA5 dataset (RMSE: 5.59~~; Yanbu: 3.77 °C) than ATM.STA run (Jeddah: 4.98 °C~~), which is probably because ERA5 uses a lower resolution grid and is unable to capture the~~ ; Yanbu: 4.29 °C) by about 1 °C and 0.5 °C, respectively. However, the difference of T2 in ~~the coastal city. This is one of~~

~~the advantages when employing regional simulations using higher resolution.~~ Mecca, which is located 70-km from the sea, is negligible (0.05 °C) between CPL and ATM.STA runs throughout the simulation. It should be mentioned that both the present simulations and ERA5 ~~dataset~~ reported a T2 that is ~~4.5~~2.8 °C ~~lower than observed~~ cooler than the observed record-high T2 in Mecca on June 2nd~~, though the heat wave events in the other cities are still captured. 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. It can be also seen in the results that taking into account ocean–atmosphere coupling can improve the simulation of~~. This under-estimation is comparable with the RMSE of the daily high T2 ~~in the CPL run. In Fig. 6, the CPL run can better reproduce the evolution of the T2 compare to ATM.STA run during the 30-day simulation: the CPL run better captures the daily high /low temperature in Yanbu and Jeddah (RMSE: 2.69 and 2.81 °C) than ATM.STA run (RMSE: 3.04 and 3.28 °C). However, the difference of T2~~ in Mecca ~~is negligible (0.05~~(2.25 °C ~~) between CPL and ATM. STA runs, shown in Fig. ??. 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.~~ in CPL run).

~~The mean and standard deviation of the surface air temperature (T2) at three major cities near the eastern shore of Red Sea (Jeddah, Mecca, Yanbu) as resulting from CPL and ATM.STA runs. Both daily high and low T2 are presented. The ATM.DYN run results are similar with the CPL run results and thus are not shown. The T2 data in all simulations are not used if they are missing in NCDC ground observation.~~

The simulation error of T2 also oscillates diurnally in the present simulations. To demonstrate the diurnal variation of the simulation error quantitatively, the mean ~~bias and RMSE~~ biases and RMSEs of T2 between the simulations (i.e., ATM.STA, ATM.DYN, and CPL) and ERA5 data are shown in Fig. 7. To ~~highlight the air–sea interactions in the simulations~~validate the coupled ocean–atmosphere model, only the temperature over the Red Sea is compared. It can be seen in Fig. 7 that the ATM.STA run using the static SST can still capture the T2 patterns in the first week, but it under-predicts T2 by ~~2.5~~about 2 °C after 20 days because of ignoring the SST evolution. On the other hand, CPL run has ~~much~~ smaller bias (~~-0.49~~-0.60 °C) and ~~root mean square error (1.46~~RMSE (1.28 °C) compared with those in ATM.STA run (bias: ~~-1.34~~-1.19 °C; RMSE: ~~2.04~~1.71 °C) during the 30-day simulation as the SST evolution is considered. The ATM.DYN run ~~uses the prescribed SST and its results are consistent with those~~ also has smaller error than ATM.STA and its error is consistent with that in CPL run (bias: ~~-0.58~~ -0.72 °C; RMSE: ~~1.40~~ 1.31 °C), indicating that the coupled ~~model captures the SST revolution. The bias and RMSE~~ simulation is comparable to the stand-alone atmosphere simulation driven by 3-hourly reanalysis SST. The biases and RMSEs of T2 in the present work are similar to those in the benchmark WRF-ARW simulations ~~(Xu et al., 2009; Zhang et al., 2013a)~~ (Xu et al., 2009; Zhang et al., 2013a; Imran et al., 2018). The differences of the mean ~~bias and RMSE~~ biases and RMSEs between the simulations and ERA5 data are also plotted to demonstrate the ~~improvement~~ evolution of the CPL ~~run over~~ errors compared with ATM.STA and ATM.DYN runs. It can be seen that the CPL run ~~captures improved T2 patterns in both~~ has smaller bias and RMSE than the ATM.STA run throughout the entire simulation. The bias and RMSE between CPL run and ATM.DYN runs are ~~consistent within~~ within about 0.5 °C. This demonstrates the capability of the coupled model ~~in~~ for performing realistic regional coupled ocean–atmosphere simulations.

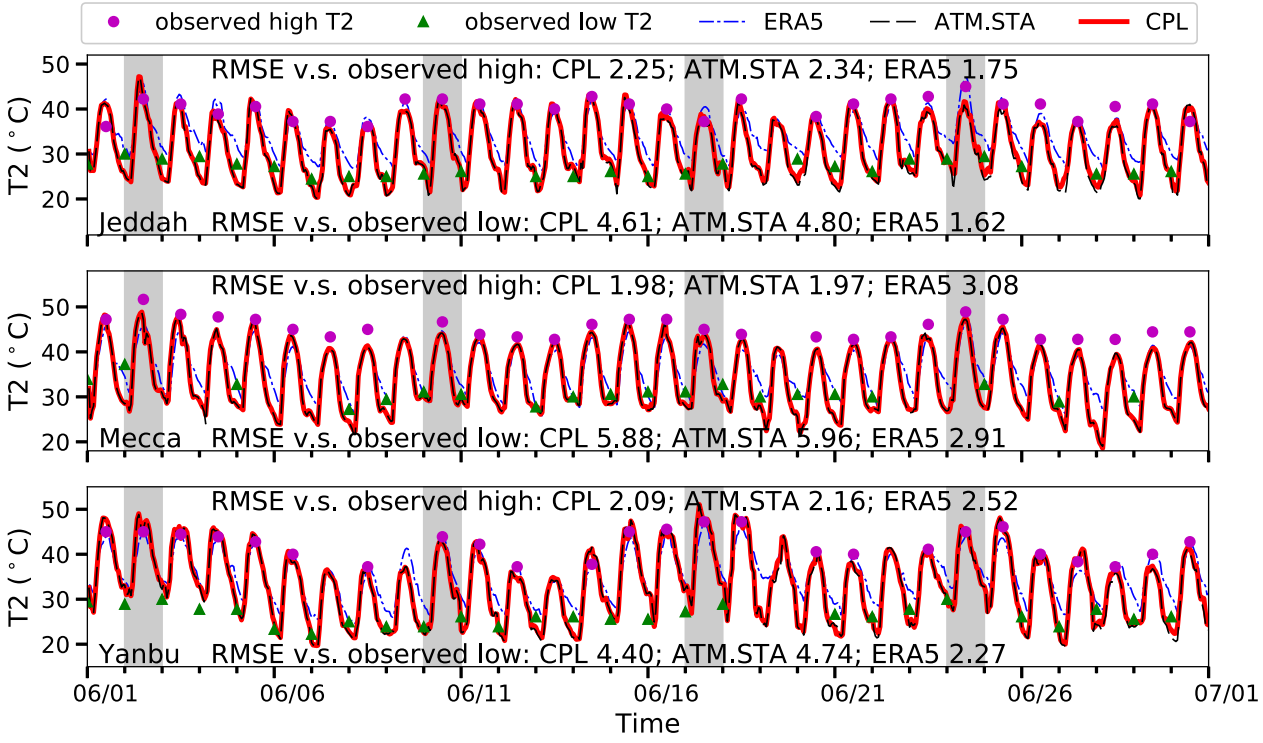


Figure 6. Temporal variation the ~~surface~~ 2-m air temperature at three major cities near the eastern shore of Red Sea (Jeddah, Mecca, Yanbu) as resulting from CPL and ATM.STA runs. The ~~ATM.DYN run results are similar with the CPL run results and thus are not shown. The~~ temperature data are compared with the time series in ERA5 ~~dataset~~ and daily high/low temperature in the NOAA national data center dataset. Note that some ~~surface~~ 2-m air temperature data gaps exist in the NCDC ground observation dataset. Four representative heat events are highlighted in this figure.

4.2 Sea Surface Temperature

The simulated SST patterns are compared to the validation data to demonstrate that the coupled model can capture the ~~performance of the coupled model in capturing the~~ ocean surface state. The ~~daily SST fields~~ SST field snapshots from CPL run on June 2nd and 24th are shown in Fig. 8(I) and Fig. 8(VI). To validate the CPL run results, the SST fields obtained in OCN.DYN runs are shown in Fig. 8(II) and 8(VII), and the GHRSST fields are shown in Fig. 8(III) and 8(VIII). The SST obtained in the model at 0000 UTC (about 3 A.M. local time in the Red Sea region) is presented because the GHRSST is produced with nighttime SST data (Roberts-Jones et al., 2012). It can be seen that both OCN.DYN and CPL runs are able to reproduce the SST patterns reasonably well in comparison with ~~the satellite observations~~GHRSST for both snapshots. Though the CPL run uses the surface forcing fields with a higher resolution, the SST patterns obtained in both simulations are very similar after two days. On June 24th, the SST patterns in both runs are less similar, but both simulation results are still ~~consistent~~ comparable with GHRSST (RMSE $< 1°C$). Both simulations under-estimate the SST in the northern Red Sea ~~. The CPL run~~

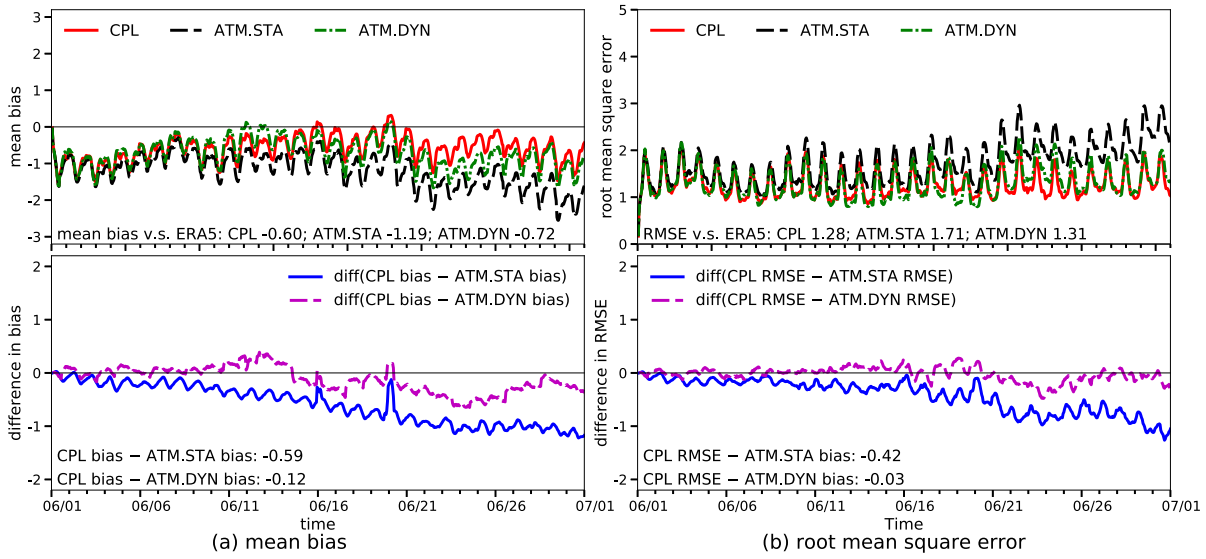


Figure 7. The bias and root mean square error (RMSE) between the ~~surface~~ 2-m air temperature obtained by the simulations (i.e., ATM.STA, ATM.CPL, and CPL) in comparison with ERA5 data. Only the errors over the Red Sea are considered. The differences between the simulation errors from CPL run and stand-alone WRF simulations are presented below the mean bias and the ~~root mean square error~~RMSE. The initial time is 0000 UTC Jun 01 2012 for all simulations.

and slightly over-estimates the SST in the central and southern Red Sea on June 24th~~, while the OCN. DYN run under-estimates the SST in the central Red Sea.~~ .

To quantitatively compare the errors in SST~~results~~, the time history of the SST in the simulations (i.e., OCN.DYN and CPL) and validation ~~datasets~~ data (i.e., GHRSST and HYCOM~~data~~/NCODA) are shown in Fig. 9. The mean bias and RMSE between simulation results and validation ~~datasets~~ data are also plotted. ~~Again, only the errors between daily SST fields are presented because both observational datasets only provide daily data. It can be seen~~ In Fig. 9(a) the snapshots of the simulated SST are compared with available HYCOM/NCODA data, in Fig. 9~~that the bias and RMSE of SST in CPL run (bias: -0.26 °C; RMSE: 0.74 °C) is smaller than that of T2 (bias: -0.47 °C; RMSE: 1.42 °C) shown in Fig. 7.~~ (b) the snapshots of SST at 0000 UTC (about 3 A.M. local time in the Red Sea region) are compared with GHRSST. Generally, the OCN.DYN and CPL runs have a similar range of error compared to both validation datasets ~~, which shows the skill of the coupled model in simulating the ocean SST.Compared with the HYCOMdataset, the bias of~~ in the 30-day simulations. The simulation results are compared with HYCOM/NCODA data to show the increase of RMSE in Fig. 9(a). Compared with HYCOM/NCODA, the mean differences between CPL and OCN.DYN runs are small (CPL: ~~-0.12~~0.10 °C; OCN.DYN: ~~-0.04~~0.03 °C)~~before June 10th. After June 11th, the CPL run slightly over-estimated the SST (0.37 °C), but the OCN.DYN run slightly under-estimated it (-0.05 °C). In addition, the RMSEs of both simulations increase~~ . The RMSE increases in the first ~~10 days, but the increase is not significant after that~~week, but does not grow after it. On the other hand, when comparing with the GHRSST, the initial SST patterns in both runs are cooler by about 0.8 °C. This is because ~~the HYCOMdata~~ our models are initialized by using HYCOM/NCODA,

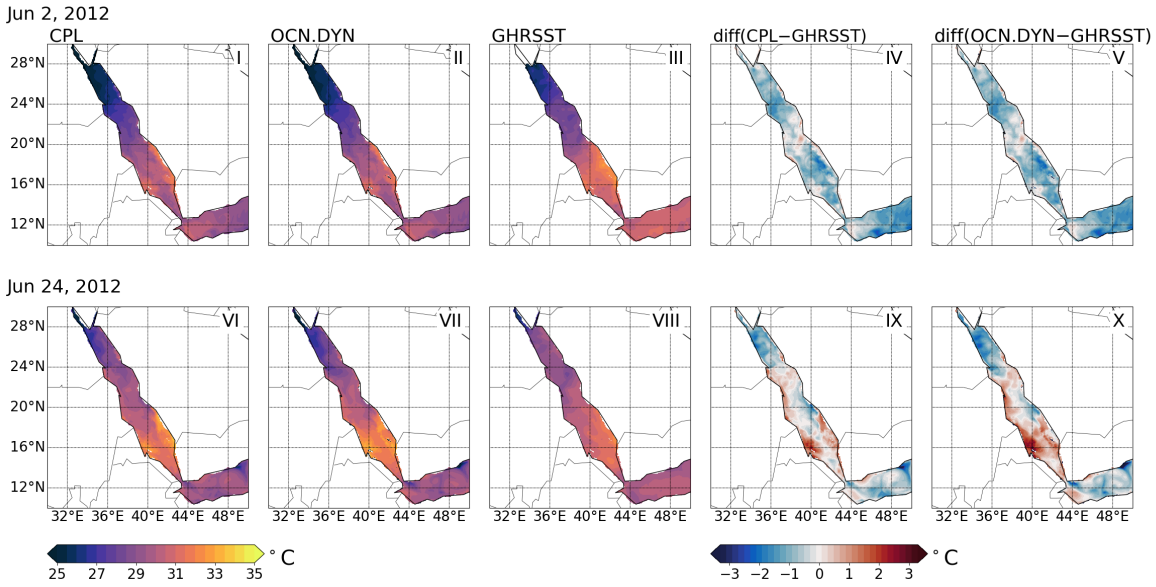


Figure 8. The daily SST patterns obtained by OCN.DYN and CPL runs, and GHRSST ~~dataset~~data. The corresponding differences between the simulations and the GHRSST ~~dataset~~ are also plotted. Two snapshots are selected: (1) 0000 UTC Jun 02 2012; (2) 0000 UTC Jun 24 2012. The simulation initial time is 0000 UTC Jun 01 2012 for both snapshots.

and the temperature in the topmost model level is cooler than ~~GHRSST at the start of the simulation~~the estimated foundation SST reported by GHRSST. After the first 10 days, the difference between GHRSST data and HYCOM/NCODA decreases, and likewise the difference between the simulation results and GHRSST also decreases. ~~Before June 10th, both CPL and ATM.STA runs under-estimated the SST (CPL : -0.73~~It should be noted that the SST simulated by the CPL run has smaller error (bias: -0.57 °C; RMSE: 0.69 °C) compared with OCN.DYN (bias: -0.66 °C~~). It should be noted that the mean SST in CPL run (-0.01~~; RMSE: 0.76 °C) ~~is closer to GHRSST than OCN.DYN (-0.34~~by about 0.1 °C ~~) after June 11th.~~ when validated using GHRSST. This indicates the coupled model can adequately simulate the SST evolution compared with the uncoupled model forced by ERA5 reanalysis data.

4.3 Surface Heat Fluxes

The ~~surface heat budget strongly influences the forecast of the surface temperature fields in the simulations. Here we evaluate the performance of the coupled modelin capturing~~ atmosphere surface heat flux drives the oceanic component in the coupled model, hence we validate the heat fluxes ~~,~~ in the coupled model as compared to the stand-alone simulations. ~~The results are also compared to the~~ Both the turbulent heat fluxes and the net downward heat fluxes are compared to MERRA-2 ~~dataset~~ and their differences are plotted. To validate the coupled ocean–atmosphere model, we only compare the heat fluxes over the sea.

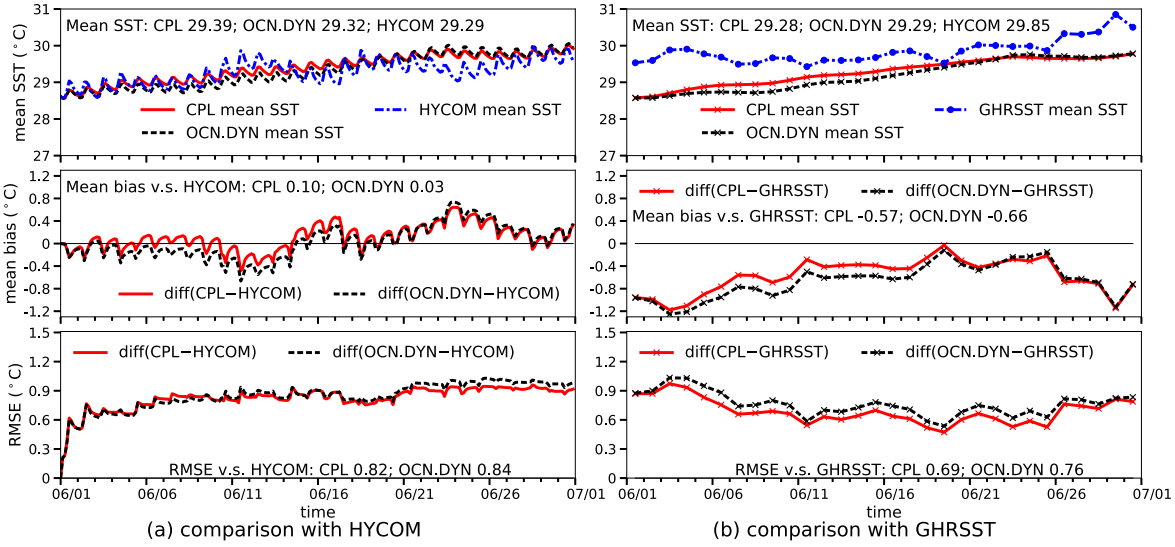


Figure 9. The bias and ~~mean-root-square-error~~ RMSE between the ~~daily~~ SST ~~as resulting~~ from the simulations (i.e., OCN.DYN and CPL) in comparison with the ~~observational dataset~~validation data. Panel (a) shows the comparison with HYCOM~~dataset~~ /NCODA data and Panel (b) shows the comparison with GHRSST~~dataset~~. The initial time is 0000 UTC Jun 01 2012 for all simulations.

The turbulent heat fluxes ~~(THF), including the latent heat~~ (THF; sum of latent and sensible heat ~~,~~ fluxes) and their differences with the validation ~~dataset~~ data are shown in Fig. ~~A1. The snapshots of the turbulent fluxes in the heat wave events on June 2nd and 24th are presented.~~ 10. It can be seen ~~that all simulations reproduce the turbulent heat fluxes reasonably well in comparison with the~~ in Fig. 10 that both CPL and ATM.STA runs capture the mean THF over the Red Sea compared with MERRA-2 ~~dataset.On June 2nd,~~ (CPL: 119.4 W/m^2; ATM.STA: 103.4 W/m^2; MERRA-2: 115.6 W/m^2). For the first two weeks, the mean THFs obtained in CPL and ATM.STA in Fig. 10 are overlapping and all simulations exhibit similar THF patterns ~~since they have the same initial conditions and air–sea interactions do not significantly impact the THF within two days. On the other hand, for the heat wave event on June 24th, CPL and ATM.DYN runs exhibit more latent heat fluxes coming out of the ocean (157 and 131~~because they are initialized in the same way and the SST fields are similar (see the snapshots comparison in the Appendix). After the second week, the CPL run has smaller error (bias: -1.8 W/m^2~~) than that in~~ ; RMSE: 69.9 W/m^2) compared with the ATM.STA run (~~115~~bias: -25.7 W/m^2~~). The mean biases in ATM.STA, ATM.DYN, and CPL runs are -9.8 w~~; RMSE: 76.4 W/~~m2, 5.9 w/m2, and 31.8 w/m2, respectively.~~ m^2). This is because the SST ~~fields in stand-alone WRF runs are cooler compared with CPL~~ is updated in the CPL run and is warmer compared with ATM.STA run. When forced by ~~cooler~~ a warmer SST, the evaporation ~~decreases~~ increases (also see the Appendix) and thus the latent heat ~~is smaller. Compared with the latent heat, the sensible heat in the Red Sea region is much smaller in all simulations (10~~fluxes increases. On the other hand, the THFs in the CPL run are comparable with the ATM.DYN run during the 30-day run (bias: 1.9 W/m^2)~~. It should be noted that the MERRA-2 dataset has unrealistically large sensible heat in the coastal regions because its resolution is not adequate to resolve~~

the coastline in the Red Sea region (Gelaro et al., 2017), showing the SKRIPS can capture the THFs over the Red Sea in the coupled simulation.

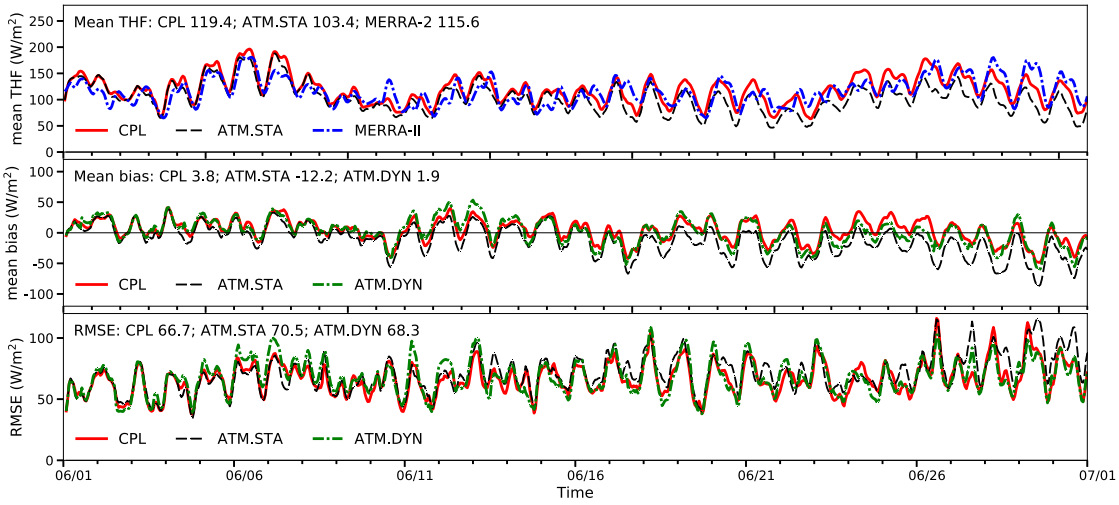


Figure 10. The turbulent heat fluxes out of the sea obtained in CPL run, MERRA-2 data, and their difference (CPL – MERRA-2). The difference between ATM.STA and ATM.DYN runs in comparison with the MERRA-2data (i. e., ATM.STA – MERRA-2, ATM.DYN – MERRA-2) are also presented. Two snapshots are selected: (1) 1200 UTC Jun 02 2012; (2) 1200 UTC Jun 24 2012. The simulation initial time is 0000 UTC Jun 01 2012 for both snapshotstop panel shows the mean THF; the middle panel shows the mean bias; the bottom panel shows the RMSE. Only the hourly heat fluxes over the sea is are shown to highlight validate the air–sea interactionscoupled model.

The net downward shortwave and longwave heat fluxesheat fluxes (sum of latent heat, sensible heat, shortwave radiation fluxes, and longwave radiation fluxes) are shown in Fig. A211. Again, all simulations reproduce the shortwave and longwave radiation fluxes reasonably well. For the shortwave heat flux, all simulations show similar patterns on both June 2nd and 24th as the air–sea interactions do not significantly impact the solar radiation. However, compared with ATM.STA run, there is a small improvement in the CPL (2.19for the first two weeks, the heat fluxes obtained in CPL and ATM.STA runs are overlapping and all simulations exhibit similar heat flux patterns because they are initialized in the same way the SST fields are similar (see the snapshots comparison in the Appendix). After the second week, the CPL run has slightly smaller error (bias: 11.2 W/m^2) and ATM.DYN (1.27; RMSE: 84.4 W/m^2) compared with the ATM.STA simulation (bias: 36.5 W/m^2) runs on June 24th. This is because these two simulations are driven by realistic SST and thus can capture longwave radiation according to the bulk formula. The total ; RMSE: 94.3 W/m^2). It should be noted that the mean bias and RMSE of the net downward heat fluxes , which is the sum of the results in Figs. A1 and A2, can be as high as a few hundred W/m^2 or 40% compared with MERRA-2. This is because WRF over-estimated the shortwave radiations at daytime (detailed comparisons are shown in Fig. A3. It can be seen that the present simulations over-estimated the total downward heat fluxes (CPL : 646the Appendix).

However, the coupled model still captures the mean and standard deviation of the heat flux compared with MERRA-2 data (CPL mean: 110.6 W/m^2~~; ATM.STA: 674~~, standard deviation: 350.7 W/m^2; ~~ATM.DYN: 663~~MERRA-2 mean 104.7 W/m^2~~)~~ ~~for both heat wave events compared with MERRA-2 dataset (495~~, standard deviation 342.3 W/m^2~~)~~, ~~especially in the central Red Sea, the southern Red Sea and the coastal regions. In the central and southern Red Sea, the~~ . The over-estimation ~~is~~ of shortwave radiation by RRTMG scheme is also reported in other validation tests in the literature under all-sky conditions due to the ~~discrepancies in shortwave solar radiation. To improve the forecast of shortwave radiation, a better understanding of the cloud and aerosol in~~ uncertainty of cloud or aerosol (Zempila et al., 2016; Imran et al., 2018). Although the surface heat flux is slightly over-estimated at daytime, the SST over the Red Sea ~~region is required. In the coastal region, the discrepancy is because MERRA-2 data are only available on a lower resolution grid and do not resolve heat fluxes in the coastal regions. It should be noted that ATM.STA run has the largest discrepancy on June 24th when using a constant SST field. Overall, the present CPL simulations are capable of well capturing all the components of the surface heat fluxes during the heat wave events.~~ is not over-estimated (shown in Section 4.2).

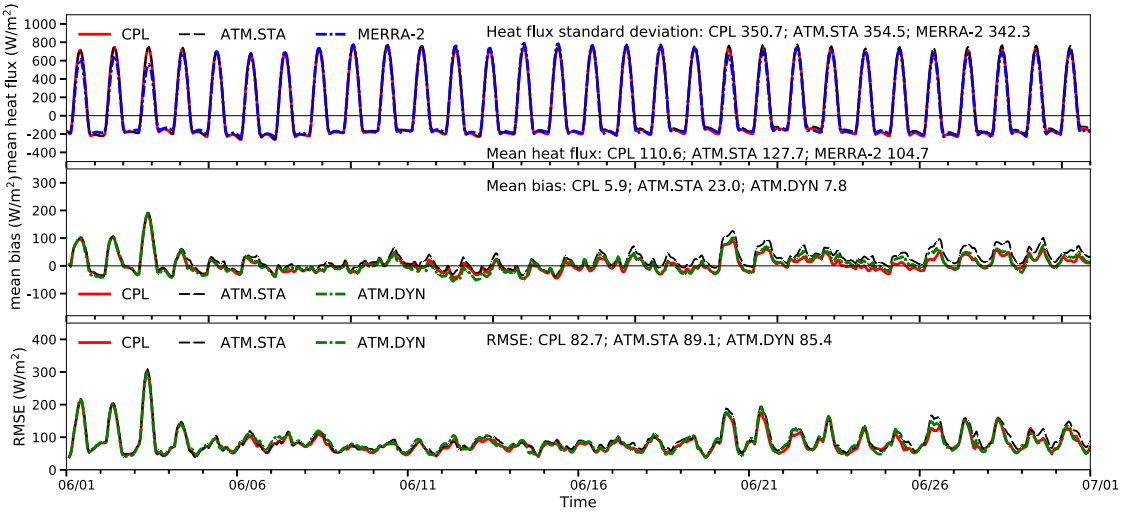


Figure 11. The ~~net downward shortwave and longwave~~ total surface heat fluxes into the sea obtained in CPL ~~run, MERRA-2 data,~~ and ~~their difference (CPL−MERRA-2). The difference between~~ ATM.STA ~~and ATM.DYN~~ runs in comparison with ~~the~~ MERRA-2~~data (i. e., ATM.STA−MERRA-2, ATM.DYN−MERRA-2) are also presented. Two snapshots are selected: (1) 1200 UTC Jun 02 2012; (2) 1200 UTC Jun 24 2012.~~ The ~~simulation initial time is 0000 UTC Jun 01 2012 for both snapshots~~top panel shows the mean surface heat flux; the middle panel shows the mean bias; the bottom panel shows the RMSE. Only the heat fluxes over the sea ~~is~~ are shown to ~~highlight~~ validate the ~~air–sea interactions~~coupled model.

~~Comparison of the total downward heat fluxes obtained in CPL run, MERRA-2 data, and their difference (CPL−MERRA-2). The difference between ATM.STA and ATM.DYN with the ERA5 data (i.e., ATM.STA−MERRA-2, ATM.DYN−MERRA-2) are also presented. Two snapshots are selected: (1) 1200 UTC Jun 02 2012; (2) 1200 UTC Jun 24 2012. The simulation~~

~~initial time is 0000 UTC Jun 01 2012 for both snapshots. Only the heat fluxes over the sea is shown to highlight the air–sea interactions.~~

4.4 ~~Surface~~ 10-m Wind~~and Evaporation~~

To evaluate the simulation of the surface momentum ~~and freshwater fluxes~~ by the coupled model, the ~~surface wind and evaporation~~ 10-m wind patterns obtained from ATM.STA, ATM.DYN, and CPL runs are presented. The MERRA-2 data ~~is~~ are used to validate the simulation results.

The simulated ~~surface~~ 10-m wind velocity fields are shown in Fig. 12. The RMSE of the wind velocity magnitude between the CPL run and MERRA-2 data is ~~2.17~~2.23 m/s when using the selected WRF physics schemes presented in Section 3. On June 2nd, high-speed wind is observed in the northern and central Red Sea, and ~~the CPL run successfully captures the small-scale~~ both the CPL and ATM.STA runs capture the features of wind speed patterns. On June 24th, the ~~differences between the simulations are larger than those on June 2nd, especially~~ high-speed wind is observed in the central Red Sea and ~~the southern Arabian Peninsula. It should be mentioned that although the SST in the~~ is also captured by both CPL and ATM.STA ~~run is lower than the CPL run, the RMSE in the wind velocity magnitude is small than 1 m/s (June 2nd: 0.15 m/s; June 24th: 0.74 m/s).~~

~~The magnitude and direction of the surface wind obtained in the CPL run, the MERRA-2 data, and their difference (CPL−MERRA-2). The difference between ATM.STA and ATM.DYN with the MERRA-2 data (i.e., ATM.STA−MERRA-2, ATM.DYN−MERRA-2) are also presented. Two snapshots are selected: (1) 1200 UTC Jun 02 2012; (2) 1200 UTC Jun 24 2012.~~

~~The surface evaporation results are shown in Fig. B1. All simulations reproduce the overall evaporation patterns in the Red Sea . The CPL run is able to capture the relatively high evaporation in the northern Red Sea and~~ runs. The mean 10-m wind speed over the Red Sea in the ~~relatively low evaporation in the southern Red Sea in both snapshots,~~ CPL and ATM.STA runs during the 30-day simulation are shown in Fig. ~~B1(I) and B1(VI). Again, all simulation results are consistent on June 2nd because they are driven by the same initial condition. However, after 24 days, the CPL run agrees better with MERRA-2 dataset (bias: 4 cm~~13. The mean error of CPL run (mean bias: -0.23 m/~~years~~; RMSE: ~~64 cm~~2.38 m/~~year)~~ s) is slightly smaller than the ATM.STA run (~~bias: -34 cm~~mean bias: -0.34 m/~~years~~; RMSE: ~~69 cm~~2.43 m/~~year) by better reproducing the realistic ocean–atmosphere coupling. Although the CPLrun results are consistent with that of the ATM.DYN run, the coupled model over-estimates the evaporation in the southern Red Sea. This is because the CPL run over-estimated the SST than the~~ s) by about 0.1 m/s. Although CPL, ATM.STA, ATM.DYN ~~run, shown in Fig. 8(IX). Since there is no precipitation in three major cities (Mecca, Jeddah, Yanbu) near the eastern shore of~~ runs have different SST as the atmospheric boundary condition, the ~~Red Sea during the month according to NCDC climate data, the precipitation results are not shown~~10-m wind speed fields obtained in the simulations are all consistent with MERRA-2 data. The comparison shows the SKRIPS is capable of simulating the surface wind speed over the Red Sea in the coupled simulation.

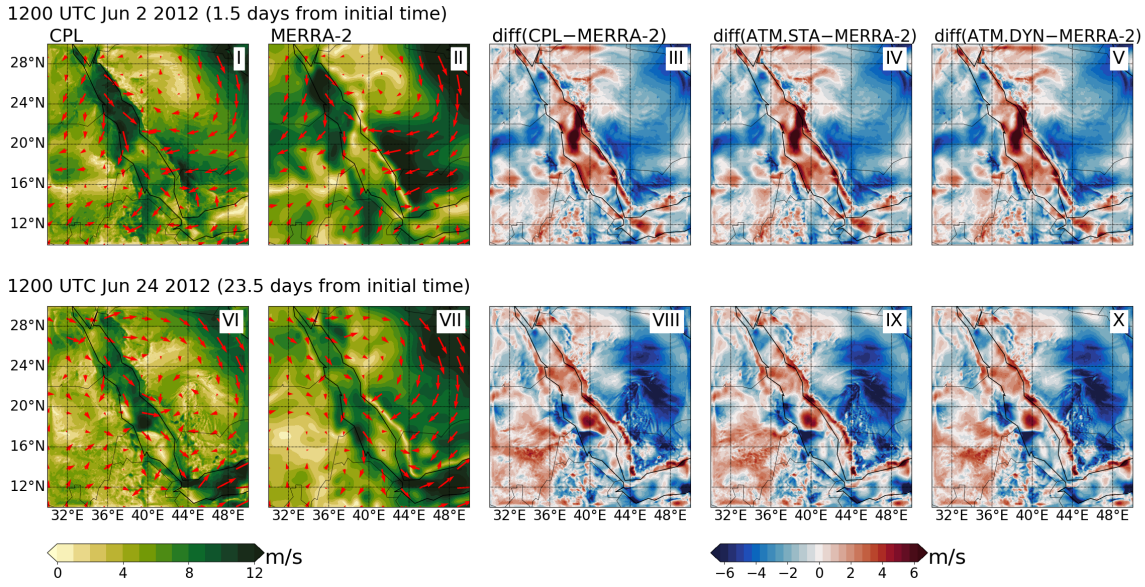


Figure 12. The ~~surface evaporation patterns~~ magnitude and direction of the 10-m wind obtained in the CPL run, the MERRA-2 data, and their difference (CPL−MERRA-2). The ~~difference~~ differences between ATM.STA and ATM.DYN with ~~the~~ MERRA-2 ~~data~~ (i.e., ATM.STA−MERRA-2, ATM.DYN−MERRA-2) are also presented. Two snapshots are selected: (1) 1200 UTC Jun 02 2012; (2) 1200 UTC Jun 24 2012.~~Only the evaporations over the sea is shown to highlight the air–sea interactions.~~

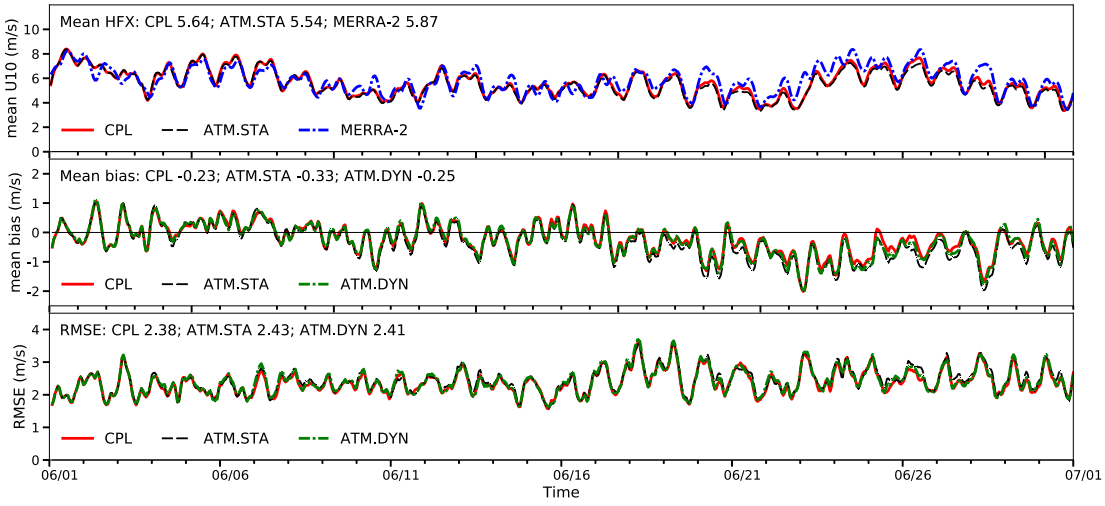


Figure 13. The magnitude of the 10-m wind obtained in CPL and ATM.STA runs in comparison with MERRA-2. The top panel shows the mean 10-m wind; the middle panel shows the mean bias; the bottom panel shows the RMSE. Only the hourly-averaged surface wind fields over the sea are shown to validate the coupled model.

5 Scalability Test

Parallel efficiency is crucial for coupled ocean–atmosphere models ~~for~~ when simulating large and complex problems. In this section, the parallel efficiency in the coupled simulations is investigated. This aims to demonstrate (1) the implemented ESMF/NUOPC driver and model interfaces ~~are able to~~ can simulate parallel cases effectively and (2) the ESMF/NUOPC coupler is not a bottleneck of the coupled simulation. The parallel speed-up of the model is investigated to evaluate its performance for a constant size problem simulated using different numbers of ~~processors~~ CPU cores (i.e. strong scaling). Additionally, the CPU time spent on ~~different parts~~ oceanic and atmospheric components of the coupled model is detailed. The parallel efficiency tests are performed on the ~~COMPAS (Center for Observations, Modeling and Prediction at Scripps) cluster in Scripps Institution of Oceanography~~ (Shaheen-II cluster in KAUST (https://www.hpc.kaust.edu.sa/). The ~~COMPAS cluster is composed of 1192 Intel 5400 and 5500 series CPUs and~~ Shaheen-II cluster is a Cray XC40 system composed of 6174 dual sockets compute nodes based on 16 cores Intel Haswell processors running at 2.3GHz. Each node has 128GB DDR4 memory running at 2300MHz. Overall the system has a total of 197,568 CPU cores (6147 nodes $\times$ 2 $\times$ 16 CPU cores) and has a theoretical peak speed of ~~12.6 TeraFlops. The cluster uses Myrinet for its high-performance network.~~ 7.2 PetaFLOPS (10^{15} floating point operations per second).

The parallel efficiency of the scalability test is $N_{p0}t_{p0}/N_{pn}t_{pn}$, where N_{p0} and N_{pn} are the ~~number of processors~~ numbers of CPU cores employed in the simulation of the baseline case and the test case, respectively; t_{p0} and t_{pn} are the CPU time spent on the baseline case and the test case, respectively. The speed-up is defined as t_{p0}/t_{pn}, which is the relative improvement of the CPU time when solving the problem. The scalability tests are performed by ~~runing 6-hour~~ running 24-hour simulations for ATM.STA, OCN.DYN, and CPL cases. There are a total number of 2.6 million atmosphere cells (256 lat$\times$256 lon$\times$40 vertical levels) and 0.4 million ocean cells (256 lat$\times$256 lon$\times$40 vertical levels, but about 84% of the domain is land and masked out). We started using N_{p0} = 32 because each compute node has 32 CPU cores. The results obtained in the scalability test of the coupled model are shown in Fig. 14. It can be seen that the parallel efficiency of the coupled code is close to 100% when employing less than 128 ~~processors~~ cores and is still as high as 70% when using 256 ~~processors~~cores. When using 256 ~~processors~~cores, there are a maximum of 20480 cells (16 lat$\times$16 lon$\times$80 vertical levels) in each ~~processor, but there are 5120 overlap cells (4 sides$\times$16 tiles per side$\times$80 vertical levels), which is 25% of the total cells. From results reported in previous literature, the parallel efficiency of the coupled model is comparable to other ocean-alone or atmosphere-alone models when having similar number of grid points per processor (Marshall et al., 1997; Zhang et al., 2013b).~~ core. The decrease in parallel efficiency results from the increase of communication time, load imbalance, and I/O (read and write) operation per ~~processor~~CPU core (Christidis, 2015). ~~It is noted in Fig. 14 that the parallel efficiency fluctuates when using 8 to 32 processors. This may be because of the fluctuation of the communication time, load imbalance, and I/O operations. The fluctuation of the CPU time can also be seen in the speed-up curve, but at smaller magnitude~~From results reported in the literature, the parallel efficiency of the coupled model is comparable to other ocean-alone or atmosphere-alone models when having similar number of grid points per CPU core (Marshall et al., 1997; Zhang et al., 2013b).

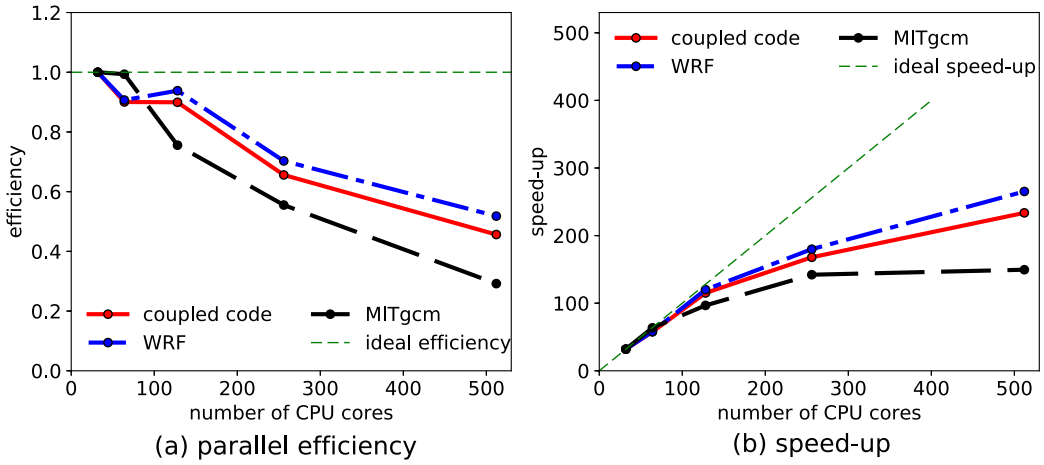


Figure 14. The parallel efficiency test of the coupled model in the Red Sea region. The test cases employ up to ~~256~~ 512 CPU cores. The simulation ~~with the smallest case~~ using 32 CPU cores is regarded as ~~base~~ the baseline case when computing the speed-up. Tests are performed on the ~~COMPAS~~ Shaheen-II cluster in ~~Scripps Institution of Oceanography~~KAUST.

The CPU time spent on ~~coupled and stand-alone runs~~ different components of the coupled run is shown in Table. 4. The time spent on the ~~coupler is estimated~~ ESMF coupler is obtained by subtracting the time spent on ~~stand-alone simulations~~ oceanic and atmospheric components from the coupled run. The most time-consuming process is the atmospheric model integration, which accounts for ~~76% to 93~~85% to 95% of the total costs. The ocean model integration is the second most time-consuming process, which is ~~7% to 14~~5% to 11% of the total computational costs. The atmospheric model is much more time-consuming because it solves the entire computational domain, while the ocean model only solves the Red Sea (16% of the domain). 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~~. If a purely marine region is selected in an ideal case, the cost of ocean and atmosphere models would be more equal compared with the Red Sea case. The coupling process takes less than ~~5~~3% of the total costs ~~when using fewer than 128 processors (40960 grid points per processor). However, when using 256 processors (20480 grid points per processor),~~ in the CPL runs using different numbers of CPU cores in this test. Although the proportion of ~~this cost increases to 10%, though the amount of~~ the coupling process in the total costs increases when using more CPU cores, the total time spent on the ~~ESMF/NUOPC coupler is similarwith using 128 processors. 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~~coupling process is similar. The CPU time spent on two uncoupled runs (i.e., ATM.STA, OCN.DYN) is also shown in Table. 4. Compared with the uncoupled simulations, the ESMF-MITgcm and ESMF-WRF interfaces do not increase the CPU time in the coupled simulation. In summary, the scalability test results suggest that the ESMF/NUOPC coupler ~~will not be~~ is not a bottleneck for using SKRIPS in coupled regional modeling studies.

Table 4. Comparison of CPU time spent on the coupled ~~run~~ and stand-alone ~~simulations~~runs. The CPU ~~times~~ time presented ~~here are~~ in the table is normalized by the time spent on the coupled run using ~~256 processors~~512 CPU cores. The CPU time spent on ~~the ESMF/NUOPC coupler is obtained by subtracting~~ two stand-alone ~~simulation time from~~ simulations are presented to show the difference between the CPL run ~~time~~and stand-alone simulations.

	$N_p =$ ~~8 16~~ 32	64	128	256	
CPL run	~~22.36~~7.27	~~11.52~~ 4.04	~~5.37~~ 2.02	~~2.89~~ 1.39	
~~ATM.STA~~ WRF in CPL run	~~20.42(916~~.88(95%)	~~10.41(903~~.82(94%)	~~4.97~~1.89(93%)	~~2.57(891~~.25(90%)	~~1.~~
~~OCN.DYN~~ MITgcm in CPL run	~~1.76(8~~0.37(5%)	0.19(5%)	~~0.93(8~~0.12(6%)	~~0.36(7~~0.11(8%)	
Coupler in CPL run	~~0.14(9~~0.02(0%)	~~0.14(14%)ESMF/NUOPC coupler~~ 0.03(1%)	~~0.17(~~0.02(1%)	~~0.18(~~0.03(2%)	
ATM.STA run	~~0.11(4%)~~6.89	~~0.07(5%)~~3.80	~~0.10(10%)~~1.84	1.22	
OCN.DYN run	0.38	0.19	0.13	0.09	

6 Conclusion and Outlook

This ~~study~~ paper describes the development of the Scripps–KAUST Regional Integrated Prediction System (SKRIPS). To build the coupled model, ~~the ESMF coupler is implemented according to NUOPC consortium~~we implement the coupler using ESMF with its NUOPC wrapper layer. The ocean model MITgcm and the atmosphere model WRF are split into initialize, run, and finalize sections, with each of them being called ~~as subroutines of~~ by the coupler as subroutines in the main function.

The ~~development activities has been focused on providing a useful coupled model for~~ coupled model is validated by using a realistic application to simulate the heat ~~wave~~ events in the Red Sea region. ~~Results~~ To validate the coupled model, results from the coupled and stand-alone simulations are compared to a wide variety of available observational and reanalysis ~~datasets, aiming to demonstrate the overall performance of the coupled model with respect to stand-alone models. The~~ data. We focus on the comparison of the surface atmospheric and oceanic variables because they are used to drive the oceanic and atmospheric components in the coupled model. From the comparison, results obtained from various configurations of coupled and stand-alone model simulations all realistically capture the ~~basic characteristics of the ocean–atmosphere state~~ surface atmospheric and oceanic variables in the Red Sea region over a 30-day simulation period. The ~~surface air temperature variations~~ 2-m air temperature in three major cities ~~are consistent with the ground observations and the heat wave events are also well captured in the CPL run. The surface flux~~ obtained in the CPL and ATM.DYN runs are comparable and better than the ATM.STA run. Other surface atmospheric fields (e.g., ~~surface~~ 2-m air temperature, surface heat fluxes, ~~surface evaporations, surface wind~~ 10-m wind speed) in the CPL run are ~~consistent with the reanalysis data~~ also comparable with the ATM.DYN run and better than the ATM.STA run over the simulation period. The SST ~~fields~~ obtained in CPL run ~~are also consistent with the satellite observation data.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~~is also better than the OCN.DYN run by about 0.1 °C compared with GHRSST.

The parallel efficiency of the coupled model is examined by simulating the Red Sea region using increasing number of ~~processors. The coupled model scales linearly for up to 128 CPUs and the parallel efficiency remains about 70% for 256 processors~~CPU cores. The parallel efficiency of the coupled model is consistent with that of the stand-alone ocean and atmosphere models when using various number of CPU cores in the test. The CPU time associated with different ~~parts~~ components of the coupled simulations is also presented, ~~suggesting good parallel efficiency in both model components and ESMFcoupler~~showing the ESMF/NUOPC driver is not a bottleneck in the computation. Hence the coupled model can be ~~applied for high-resolution~~ implemented for coupled regional modeling studies on ~~massively parallel processing supercomputers~~ supercomputers with comparable performance as that attained by uncoupled stand-alone models.

~~These preliminary results~~ The results presented here 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~~. This regional coupled ~~model can be further~~ forecasting system can be improved by developing coupled data assimilation capabilities ~~on initializing coupled forecastsfrom an assimilated analysis state~~for initializing the forecasts. In addition, the model physics and model uncertainty representation in the coupled system can be enhanced using advanced techniques, such as stochastic physics parameterizations. Future work will involve exploring these and other aspects of further developing a regional coupled modeling system ~~that is best~~ suited for forecasting and process understanding purposes.

Code and data availability. The coupled model, documentation, and tutorial cases used in this work are available at https://library.ucsd.edu/dc/collection/bb1847661c, and the source code is maintained on Github https://github.com/iurnus/scripps_kaust_model. ECMWF ERA5 data are used as the atmospheric initial and boundary conditions. The ocean model uses the assimilated HYCOM/NCODA $1/12^\circ$ global analysis data as initial and boundary conditions. To validate the simulated SST data, we use the OSTIA (Operational Sea Surface Temperature and Sea Ice Analysis) system in GHRSST (Group for High Resolution Sea Surface Temperature). The simulated 2-meter air temperature (T2) is validated against the ECMWF ERA5. The observed daily maximum and minimum temperatures from NOAA National Climate Data Center is used to validate the T2 in three major cities. Surface heat fluxes (e.g., latent heat fluxes, sensible heat fluxes, longwave and shortwave radiations), which are important for ocean–atmosphere interactions, are compared with MERRA-2 (Modern-Era Retrospective analysis for Research and Applications, version 2).

Appendix A: Snapshots of Surface Heat Fluxes

The snapshots of the THFs in the simulations at 1200 UTC June 2nd and 24th are presented. It can be seen that all simulations reproduce the THFs reasonably well in comparison with MERRA-2. On June 2nd, all simulations exhibit similar THF patterns since they have the same initial conditions and similar SST fields. On the other hand, for the heat event on June 24th, CPL and ATM.DYN runs exhibit more latent heat fluxes coming out of the ocean (170 and 153 W/m^2) than that in ATM.STA run

(138 W/m^2). The mean biases in CPL, ATM.DYN, and ATM.STA runs are 23.1 w/m^2, 5.1 w/m^2, and -9.5 w/m^2, respectively. Although the CPL run has larger bias at the snapshot, the averaged bias and RMSE in CPL run is smaller (shown in Fig. 10). Compared with the latent heat fluxes, the sensible heat fluxes in the Red Sea region are much smaller in all simulations (about 20 W/m^2). It should be noted that MERRA-2 has unrealistically large sensible heat fluxes in the coastal regions because the land points are 'contaminated' the values in the coastal region (Kara et al., 2008; Gelaro et al., 2017), and thus the heat fluxes in the coastal regions are not shown.

The net downward shortwave and longwave heat fluxes are shown in Fig. A2. Again, all simulations reproduce the shortwave and longwave radiation fluxes reasonably well. For the shortwave heat fluxes, all simulations show similar patterns on both June 2nd and 24th. The total downward heat fluxes, which is the sum of the results in Figs. A1 and A2, are shown in Fig. A3. It can be seen that the present simulations over-estimated the total downward heat fluxes on June 2nd (CPL: 580 W/m^2; ATM.STA: 590 W/m^2; ATM.DYN: 582 W/m^2) for both heat events compared with MERRA-2 (525 W/m^2), especially in the southern Red Sea because of the over-estimation of the shortwave radiation. To improve the modeling of shortwave radiation, a better understanding of the cloud and aerosol in the Red Sea region is required (Zempila et al., 2016; Imran et al., 2018). Again, the heat fluxes in the coastal regions are not shown because of the inconsistency of land-sea mask. Overall, the comparison shows the present CPL simulations are capable of capturing the surface heat fluxes into the ocean.

Appendix B: Evaporation

To examine the simulation of surface freshwater flux in the coupled model, the surface evaporation fields obtained from ATM.STA, ATM.DYN, and CPL runs are presented and validated using the MERRA-2 data.

The surface evaporation fields from CPL run are shown in Fig. B1. The MERRA-2 data and difference between CPL run and MERRA-2 are also shown to validate the CPL run. The ATM.STA and ATM.DYN simulation results are not shown, but their differences with the CPL run are also shown in Fig. B1. It can be seen in Fig. B1(III) and B1(VIII) that the CPL run reproduces the overall evaporation patterns in the Red Sea. The CPL run is able to capture the relatively high evaporation in the northern Red Sea and the relatively low evaporation in the southern Red Sea in both snapshots, shown in Fig. B1(I) and B1(VI). After 36-hours, the simulation results are close with each other (e.g., the RMSE between CPL and ATM.STA simulation is smaller than 10 cm/year). However, after 24 days, the CPL run agrees better with MERRA-2 (bias: 6 cm/year; RMSE: 59 cm/year) than the ATM.STA run (bias: -25 cm/year; RMSE: 68 cm/year). On the other hand, the CPL run results are consistent with the ATM.DYN run. This shows the CPL run can reproduce the realistic evaporation patterns over the Red Sea in the coupled ocean–atmosphere simulation. Since there is no precipitation in three major cities (Mecca, Jeddah, Yanbu) near the eastern shore of the Red Sea during the month according to NCDC climate data, the precipitation results are not shown.

Author contributions. RS worked on the coding tasks for coupling WRF with MITgcm using ESMF, wrote the code documentation, and performed the simulations for the numerical experiments. RS and ACS worked on the technical details for debugging the model and drafted

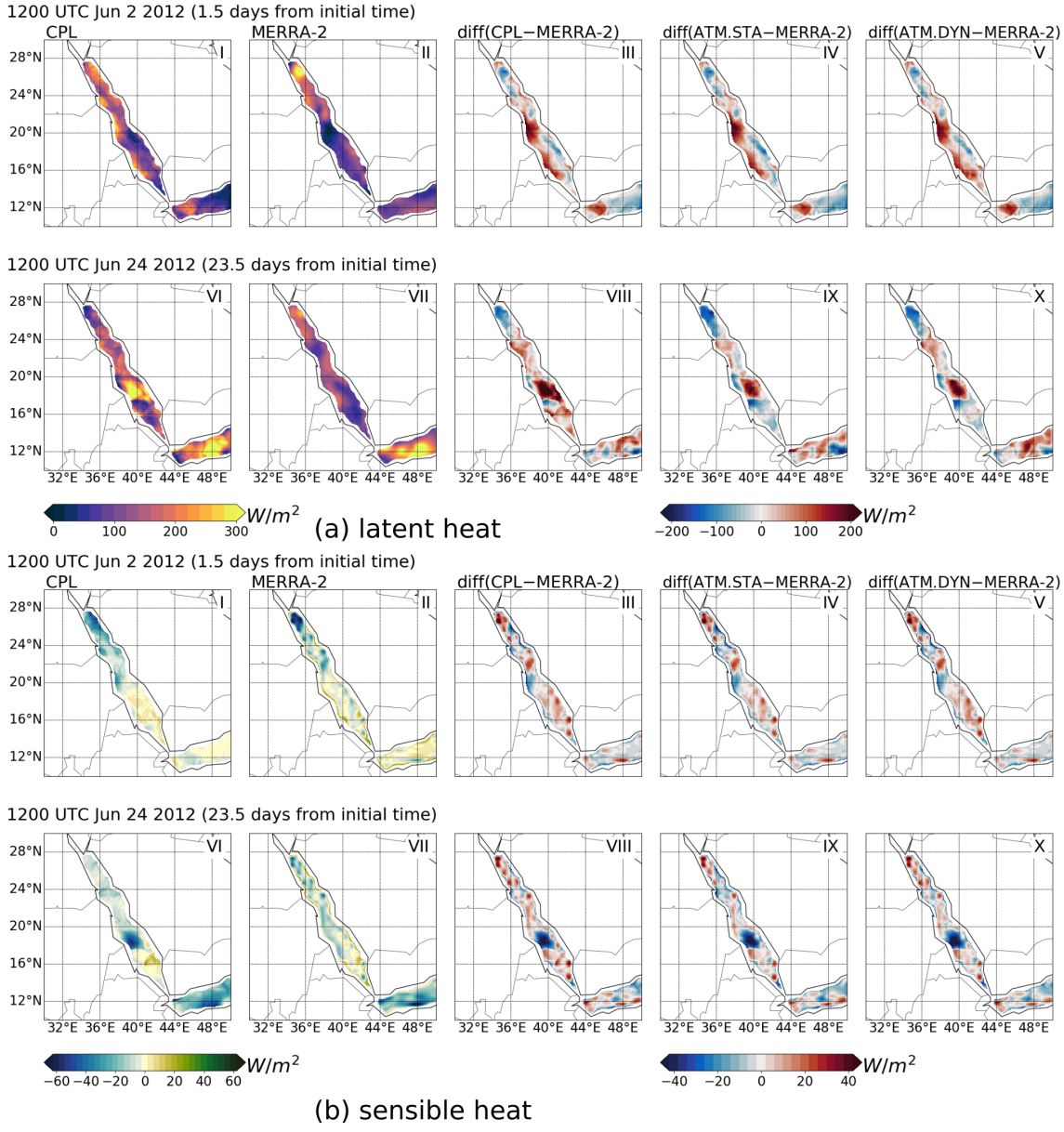
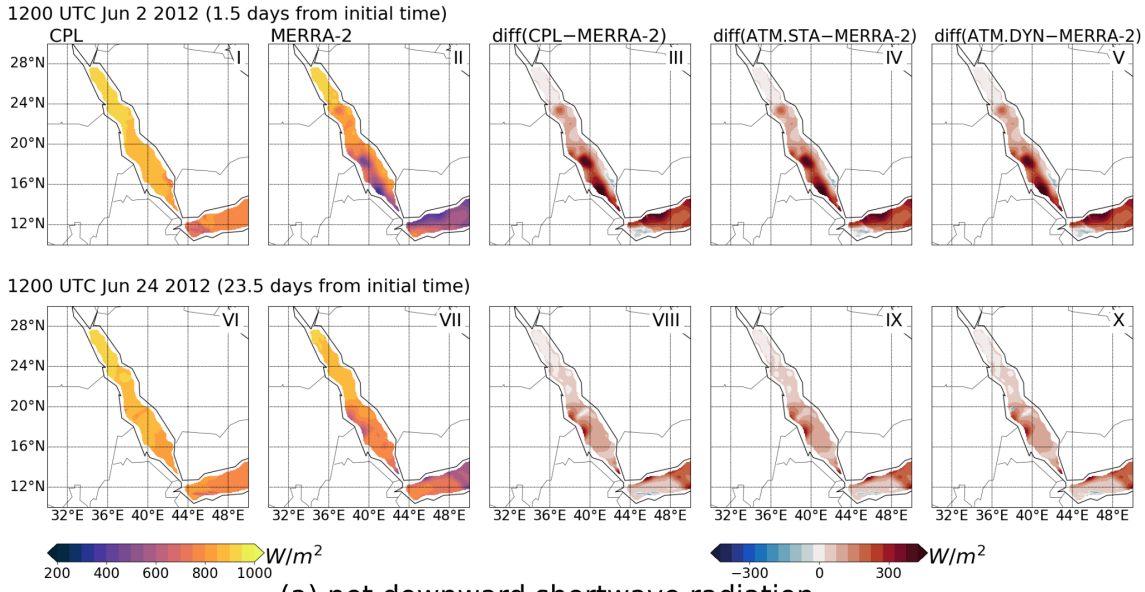
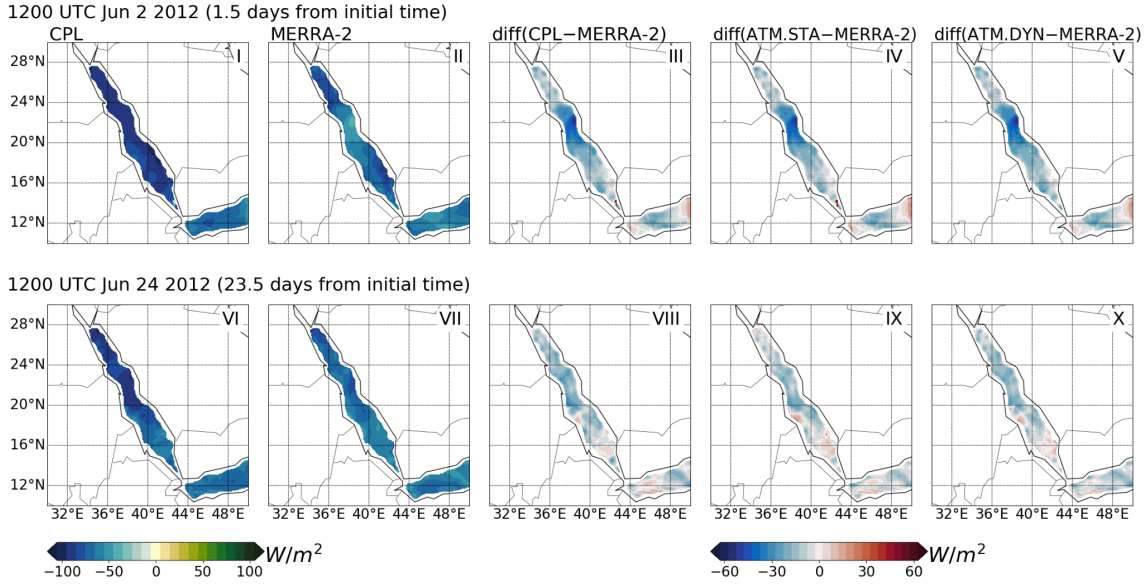


Figure A1. The turbulent heat fluxes out of the sea obtained in CPL run, MERRA-2 data, and their difference (CPL−MERRA-2). The differences between ATM.STA and ATM.DYN with MERRA-2 (i.e., ATM.STA−MERRA-2, ATM.DYN−MERRA-2) are also presented. Two snapshots are selected: (1) 1200 UTC Jun 02 2012; (2) 1200 UTC Jun 24 2012. The simulation initial time is 0000 UTC Jun 01 2012 for both snapshots. Only the heat fluxes over the sea are shown to validate the coupled model.



(a) net downward shortwave radiation



(b) net downward longwave radiation

Figure A2. The net downward shortwave and longwave heat fluxes obtained in CPL run, MERRA-2 data, and their difference (CPL−MERRA-2). The differences between ATM.STA and ATM.DYN with MERRA-2 (i.e., ATM.STA−MERRA-2, ATM.DYN−MERRA-2) are also presented. Two snapshots are selected: (1) 1200 UTC Jun 02 2012; (2) 1200 UTC Jun 24 2012. The simulation initial time is 0000 UTC Jun 01 2012 for both snapshots. Only the heat fluxes over the sea are shown to validate the coupled model.

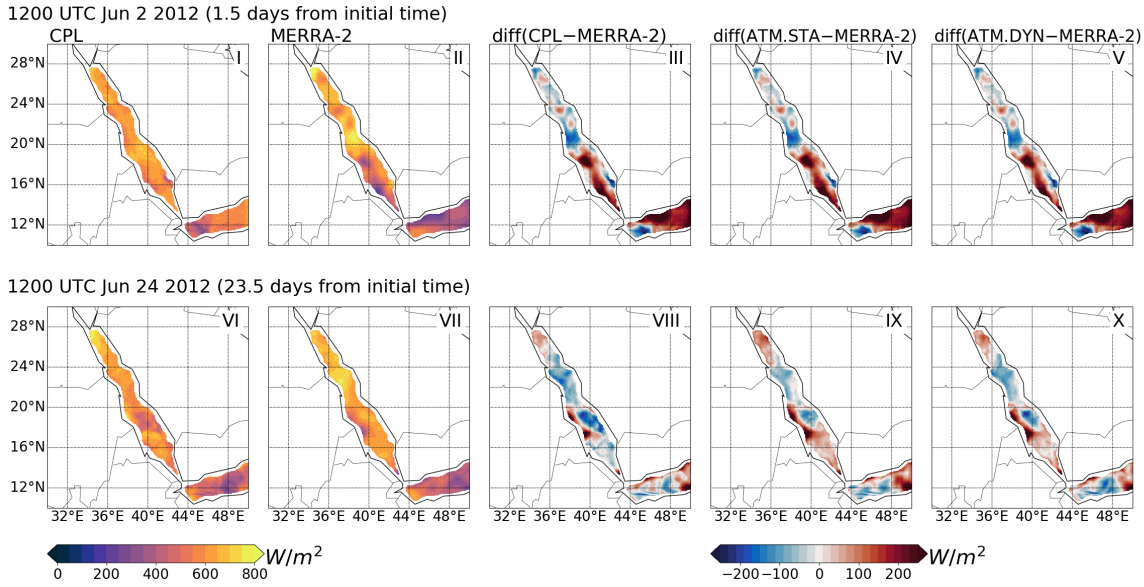


Figure A3. Comparison of the total downward heat fluxes obtained in CPL run, MERRA-2 data, and their difference (CPL−MERRA-2). The differences between ATM.STA and ATM.DYN with ERA5 (i.e., ATM.STA−MERRA-2, ATM.DYN−MERRA-2) are also presented. Two snapshots are selected: (1) 1200 UTC Jun 02 2012; (2) 1200 UTC Jun 24 2012. The simulation initial time is 0000 UTC Jun 01 2012 for both snapshots. Only the heat fluxes over the sea are shown to validate the coupled model.

the initial manuscript. All authors designed the computational framework and the numerical experiments. All authors discussed the results and contributed to the writing of the final manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We appreciate the computational resources provided by COMPAS (Center for Observations, Modeling and Prediction at Scripps) and KAUST used for this project. We gratefully acknowledge the research funding from KAUST (grant number: OSR-2-16-RPP-3268.02). We are immensely grateful to Caroline Papadopoulos for helping with installing software, testing the coupled code, and using the ~~COMPAS cluster~~HPC clusters. We appreciate Professor U.U. Turuncoglu sharing part of their ESMF/NUOPC code on GitHub which helped our code development. We wish to thank Dr. Ganesh Gopalakrishnan for setting up the stand-alone MITgcm simulation (OCN.DYN) and providing the external forcings. We thank Drs. Stephanie Dutkiewicz, Jean-Michel Campin, Chris Hill, Dimitris Menemenlis for providing their ESMF–MITgcm interface. We ~~also~~ wish to thank Dr. Peng Zhan for discussing the simulations of the Red Sea. We also thank the reviewers for their insightful review suggestions.

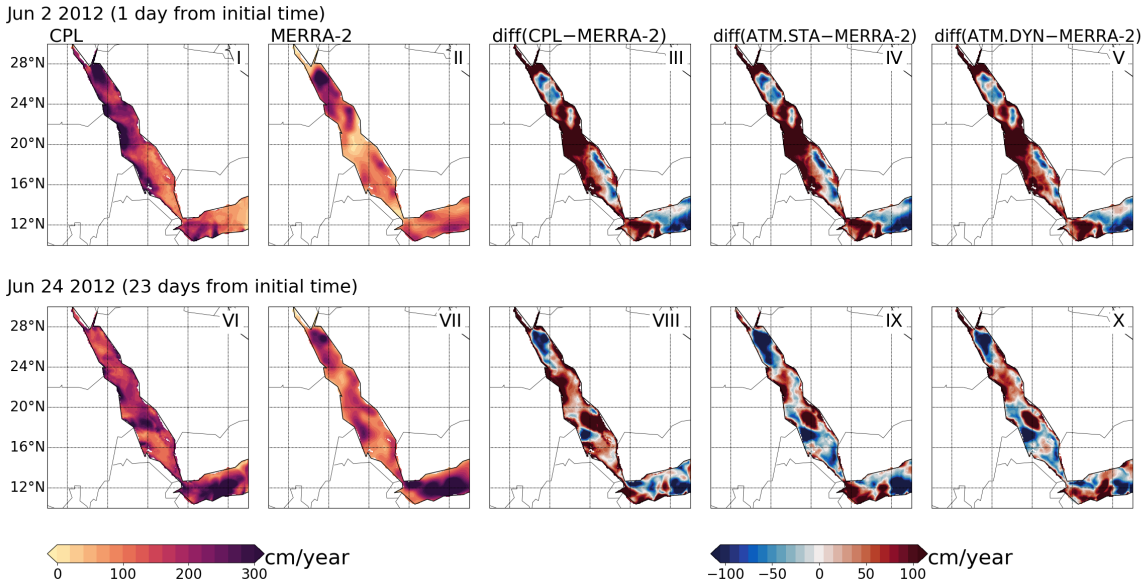


Figure B1. The surface evaporation patterns obtained in the CPL run, the MERRA-2 data, and their difference (CPL−MERRA-2). The differences between uncoupled atmosphere simulations with MERRA-2 (i.e., ATM.STA−MERRA-2, ATM.DYN−MERRA-2) are also presented. Two snapshots are selected: (1) 1200 UTC Jun 02 2012; (2) 1200 UTC Jun 24 2012. Only the evaporation over the sea is shown to validate the coupled ocean–Atmosphere model.

References

Abdou, A. E. A.: Temperature trend on Makkah, Saudi Arabia, Atmospheric and Climate Sciences, 4, 457–481, 2014.

Aldrian, E., Sein, D., Jacob, D., Gates, L. D., and Podzun, R.: Modelling Indonesian rainfall with a coupled regional model, Climate Dynamics, 25, 1–17, 2005.

Barbariol, F., Benetazzo, A., Carniel, S., and Sclavo, M.: Improving the assessment of wave energy resources by means of coupled wave-ocean numerical modeling, Renewable Energy, 60, 462–471, 2013.

Bender, M. A. and Ginis, I.: Real-case simulations of hurricane–ocean interaction using a high-resolution coupled model: effects on hurricane intensity, Monthly Weather Review, 128, 917–946, 2000.

Benjamin, S. G., Grell, G. A., Brown, J. M., Smirnova, T. G., and Bleck, R.: Mesoscale weather prediction with the RUC hybrid isentropic–terrain-following coordinate model, Monthly Weather Review, 132, 473–494, 2004.

Boé, J., Hall, A., Colas, F., McWilliams, J. C., Qu, X., Kurian, J., and Kapnick, S. B.: What shapes mesoscale wind anomalies in coastal upwelling zones?, Climate Dynamics, 36, 2037–2049, 2011.

Chen, S. S. and Curcic, M.: Ocean surface waves in Hurricane Ike (2008) and Superstorm Sandy (2012): Coupled model predictions and observations, Ocean Modelling, 103, 161–176, 2016.

Chen, S. S., Price, J. F., Zhao, W., Donelan, M. A., and Walsh, E. J.: The CBLAST-Hurricane program and the next-generation fully coupled atmosphere–wave–ocean models for hurricane research and prediction, Bulletin of the American Meteorological Society, 88, 311–318, 2007.

Christidis, Z.: Performance and Scaling of WRF on Three Different Parallel Supercomputers, in: International Conference on High Performance Computing, pp. 514–528, Springer, 2015.

Collins, N., Theurich, G., Deluca, C., Suarez, M., Trayanov, A., Balaji, V., Li, P., Yang, W., Hill, C., and Da Silva, A.: Design and implementation of components in the Earth System Modeling Framework, The International Journal of High Performance Computing Applications, 19, 341–350, 2005.

Donlon, C. J., Martin, M., Stark, J., Roberts-Jones, J., Fiedler, E., and Wimmer, W.: The operational sea surface temperature and sea ice analysis (OSTIA) system, Remote Sensing of Environment, 116, 140–158, 2012.

Doscher, R., Willén, U., Jones, C., Rutgersson, A., Meier, H. M., Hansson, U., and Graham, L. P.: The development of the regional coupled ocean-atmosphere model RCAO, Boreal Environment Research, 7, 183–192, 2002.

Evangelinos, C. and Hill, C. N.: A schema based paradigm for facile description and control of a multi-component parallel, coupled atmosphere-ocean model, in: Proceedings of the 2007 Symposium on Component and Framework Technology in High-Performance and Scientific Computing, pp. 83–92, ACM, 2007.

Fairall, C., Bradley, E. F., Hare, J., Grachev, A., and Edson, J.: Bulk parameterization of air–sea fluxes: Updates and verification for the COARE algorithm, Journal of Climate, 16, 571–591, 2003.

Fang, Y., Zhang, Y., Tang, J., and Ren, X.: A regional air-sea coupled model and its application over East Asia in the summer of 2000, Advances in Atmospheric Sciences, 27, 583–593, 2010.

Fowler, H. and Ekström, M.: Multi-model ensemble estimates of climate change impacts on UK seasonal precipitation extremes, International Journal of Climatology, 29, 385–416, 2009.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., et al.: The modern-era retrospective analysis for research and applications, version 2 (MERRA-2), Journal of Climate, 30, 5419–5454, 2017.

Gualdi, S., Somot, S., Li, L., Artale, V., Adani, M., Bellucci, A., Braun, A., Calmanti, S., Carillo, A., Dell'Aquila, A., et al.: The CIRCE simulations: regional climate change projections with realistic representation of the Mediterranean Sea, Bulletin of the American Meteorological Society, 94, 65–81, 2013.

Gustafsson, N., Nyberg, L., and Omstedt, A.: Coupling of a high-resolution atmospheric model and an ocean model for the Baltic Sea, Monthly Weather Review, 126, 2822–2846, 1998.

Hagedorn, R., Lehmann, A., and Jacob, D.: A coupled high resolution atmosphere-ocean model for the BALTEX region, Meteorologische Zeitschrift, 9, 7–20, 2000.

Harley, C. D., Randall Hughes, A., Hultgren, K. M., Miner, B. G., Sorte, C. J., Thornber, C. S., Rodriguez, L. F., Tomanek, L., and Williams, S. L.: The impacts of climate change in coastal marine systems, Ecology Letters, 9, 228–241, 2006.

He, J., He, R., and Zhang, Y.: Impacts of air-sea interactions on regional air quality predictions using WRF/Chem v3.6.1 coupled with ROMS v3.7: southeastern US example, Geoscientific Model Development Discussions, 8, 9965–10 009, 2015.

Henderson, T. and Michalakes, J.: WRF ESMF Development, in: 4th ESMF Community Meeting, Cambridge, USA, Jul 21, 2005.

Hersbach, H.: The ERA5 Atmospheric Reanalysis., in: AGU Fall Meeting Abstracts, San Francisco, USA, Dec 12-16, 2016.

Hill, C., DeLuca, C., Balaji, Suarez, M., and Silva, A.: The architecture of the Earth system modeling framework, Computing in Science & Engineering, 6, 18–28, 2004.

Hill, C. N.: Adoption and field tests of M.I.T General Circulation Model (MITgcm) with ESMF, in: 4th Annual ESMF Community Meeting, Cambridge, USA, Jul 20-21, 2005.

Hodur, R. M.: The Naval Research Laboratory's coupled ocean/atmosphere mesoscale prediction system (COAMPS), Monthly Weather Review, 125, 1414–1430, 1997.

Hong, S.-Y., Noh, Y., and Dudhia, J.: A new vertical diffusion package with an explicit treatment of entrainment processes, Monthly Weather Review, 134, 2318–2341, 2006.

Huang, B., Schopf, P. S., and Shukla, J.: Intrinsic ocean–atmosphere variability of the tropical Atlantic Ocean, Journal of Climate, 17, 2058–2077, 2004.

Iacono, M. J., Delamere, J. S., Mlawer, E. J., Shephard, M. W., Clough, S. A., and Collins, W. D.: Radiative forcing by long-lived greenhouse gases: calculations with the AER radiative transfer models, Journal of Geophysical Research: Atmospheres, 113, 2008.

Imran, H., Kala, J., Ng, A., and Muthukumaran, S.: An evaluation of the performance of a WRF multi-physics ensemble for heatwave events over the city of Melbourne in southeast Australia, Climate dynamics, 50, 2553–2586, 2018.

Kain, J. S.: The Kain–Fritsch convective parameterization: an update, Journal of Applied Meteorology, 43, 170–181, 2004.

Kara, A. B., Wallcraft, A. J., Barron, C. N., Hurlburt, H. E., and Bourassa, M.: Accuracy of 10 m winds from satellites and NWP products near land-sea boundaries, Journal of Geophysical Research: Oceans, 113, 2008.

Kharin, V. V. and Zwiers, F. W.: Changes in the extremes in an ensemble of transient climate simulations with a coupled atmosphere–ocean GCM, Journal of Climate, 13, 3760–3788, 2000.

Large, W. G. and Yeager, S. G.: Diurnal to decadal global forcing for ocean and sea-ice models: the data sets and flux climatologies, Tech. rep., NCAR Technical Note: NCAR/TN-460+STR. CGD Division of the National Center for Atmospheric Research, 2004.

Large, W. G., McWilliams, J. C., and Doney, S. C.: Oceanic vertical mixing: A review and a model with a nonlocal boundary layer parameterization, Reviews of Geophysics, 32, 363–403, 1994.

Li, D., Zou, L., and Zhou, T.: Regional air–sea coupled model simulation for two types of extreme heat in North China, Climate Dynamics, 50, 2107–2120, 2018.

Loglisci, N., Qian, M., Rachev, N., Cassardo, C., Longhetto, A., Purini, R., Trivero, P., Ferrarese, S., and Giraud, C.: Development of an atmosphere-ocean coupled model and its application over the Adriatic Sea during a severe weather event of Bora wind, Journal of Geophysical Research: Atmospheres, 109, 2004.

Marshall, J., Adcroft, A., Hill, C., Perelman, L., and Heisey, C.: A finite-volume, incompressible Navier Stokes model for studies of the ocean on parallel computers, Journal of Geophysical Research: Oceans, 102, 5753–5766, 1997.

Martin, M., Dash, P., Ignatov, A., Banzon, V., Beggs, H., Brasnett, B., Cayula, J.-F., Cummings, J., Donlon, C., Gentemann, C., et al.: Group for High Resolution Sea Surface Temperature (GHRSST) analysis fields inter-comparisons. Part 1: A GHRSST multi-product ensemble (GMPE), Deep Sea Research Part II: Topical Studies in Oceanography, 77, 21–30, 2012.

Morrison, H., Thompson, G., and Tatarskii, V.: Impact of cloud microphysics on the development of trailing stratiform precipitation in a simulated squall line: Comparison of one-and two-moment schemes, Monthly Weather Review, 137, 991–1007, 2009.

National Geophysical Data Center: 2-minute Gridded Global Relief Data (ETOPO2) v2, 2006.

Powers, J. G. and Stoelinga, M. T.: A coupled air–sea mesoscale model: Experiments in atmospheric sensitivity to marine roughness, Monthly Weather Review, 128, 208–228, 2000.

Pullen, J., Doyle, J. D., and Signell, R. P.: Two-way air–sea coupling: A study of the Adriatic, Monthly Weather Review, 134, 1465–1483, 2006.

Ricchi, A., Miglietta, M. M., Falco, P. P., Benetazzo, A., Bonaldo, D., Bergamasco, A., Sclavo, M., and Carniel, S.: On the use of a coupled ocean–atmosphere–wave model during an extreme cold air outbreak over the Adriatic Sea, Atmospheric Research, 172, 48–65, 2016.

Roberts-Jones, J., Fiedler, E. K., and Martin, M. J.: Daily, global, high-resolution SST and sea ice reanalysis for 1985–2007 using the OSTIA system, Journal of Climate, 25, 6215–6232, 2012.

Roessig, J. M., Woodley, C. M., Cech, J. J., and Hansen, L. J.: Effects of global climate change on marine and estuarine fishes and fisheries, Reviews in Fish Biology and Fisheries, 14, 251–275, 2004.

Seo, H.: Distinct influence of air–sea interactions mediated by mesoscale sea surface temperature and surface current in the Arabian Sea, Journal of Climate, 30, 8061–8080, 2017.

Seo, H., Miller, A. J., and Roads, J. O.: The Scripps Coupled Ocean–Atmosphere Regional (SCOAR) model, with applications in the eastern Pacific sector, Journal of Climate, 20, 381–402, 2007.

Seo, H., Subramanian, A. C., Miller, A. J., and Cavanaugh, N. R.: Coupled impacts of the diurnal cycle of sea surface temperature on the Madden–Julian oscillation, Journal of Climate, 27, 8422–8443, 2014.

Sitz, L., Di Sante, F., Farneti, R., Fuentes-Franco, R., Coppola, E., Mariotti, L., Reale, M., Sannino, G., Barreiro, M., Nogherotto, R., et al.: Description and evaluation of the Earth System Regional Climate Model (RegCM-ES), Journal of Advances in Modeling Earth Systems, 2017.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Wang, W., and Powers, J. G.: A description of the advanced research WRF version 2, Tech. rep., National Center For Atmospheric Research Boulder Co Mesoscale and Microscale Meteorology Div, 2005.

Somot, S., Sevault, F., Déqué, M., and Crépon, M.: 21st century climate change scenario for the Mediterranean using a coupled atmosphere–ocean regional climate model, Global and Planetary Change, 63, 112–126, 2008.

Theurich, G., DeLuca, C., Campbell, T., Liu, F., Saint, K., Vertenstein, M., Chen, J., Oehmke, R., Doyle, J., Whitcomb, T., et al.: The earth system prediction suite: toward a coordinated US modeling capability, Bulletin of the American Meteorological Society, 97, 1229–1247, 2016.

Turuncoglu, U., Giuliani, G., Elguindi, N., and Giorgi, F.: Modelling the Caspian Sea and its catchment area using a coupled regional atmosphere-ocean model (RegCM4-ROMS): model design and preliminary results, Geoscientific Model Development, 6, 283, 2013.

Turuncoglu, U. U.: Toward modular in situ visualization in Earth system models: the regional modeling system RegESM 1.1, Geoscientific Model Development, 12, 233–259, 2019.

Turuncoglu, U. U. and Sannino, G.: Validation of newly designed regional earth system model (RegESM) for Mediterranean Basin, Climate Dynamics, 48, 2919–2947, 2017.

Van Pham, T., Brauch, J., Dieterich, C., Frueh, B., and Ahrens, B.: New coupled atmosphere-ocean-ice system COSMO-CLM/NEMO: assessing air temperature sensitivity over the North and Baltic Seas, Oceanologia, 56, 167–189, 2014.

Warner, J. C., Armstrong, B., He, R., and Zambon, J. B.: Development of a coupled ocean–atmosphere–wave–sediment transport (COAWST) modeling system, Ocean Modelling, 35, 230–244, 2010.

Xie, S.-P., Miyama, T., Wang, Y., Xu, H., De Szoeke, S. P., Small, R. J. O., Richards, K. J., Mochizuki, T., and Awaji, T.: A regional ocean–atmosphere model for eastern Pacific climate: toward reducing tropical biases, Journal of Climate, 20, 1504–1522, 2007.

Xu, J., Rugg, S., Byerle, L., and Liu, Z.: Weather forecasts by the WRF-ARW model with the GSI data assimilation system in the complex terrain areas of southwest Asia, Weather and Forecasting, 24, 987–1008, 2009.

Yao, F., Hoteit, I., Pratt, L. J., Bower, A. S., Köhl, A., Gopalakrishnan, G., and Rivas, D.: Seasonal overturning circulation in the Red Sea: 2. Winter circulation, Journal of Geophysical Research: Oceans, 119, 2263–2289, 2014a.

Yao, F., Hoteit, I., Pratt, L. J., Bower, A. S., Zhai, P., Köhl, A., and Gopalakrishnan, G.: Seasonal overturning circulation in the Red Sea: 1. Model validation and summer circulation, Journal of Geophysical Research: Oceans, 119, 2238–2262, 2014b.

Zempila, M.-M., Giannaros, T. M., Bais, A., Melas, D., and Kazantzidis, A.: Evaluation of WRF shortwave radiation parameterizations in predicting Global Horizontal Irradiance in Greece, Renewable Energy, 86, 831–840, 2016.

Zhan, P., Subramanian, A. C., Yao, F., and Hoteit, I.: Eddies in the Red Sea: A statistical and dynamical study, Journal of Geophysical Research: Oceans, 119, 3909–3925, 2014.

Zhang, H., Pu, Z., and Zhang, X.: Examination of errors in near-surface temperature and wind from WRF numerical simulations in regions of complex terrain, Weather and Forecasting, 28, 893–914, 2013a.

Zhang, X., Huang, X.-Y., and Pan, N.: Development of the upgraded tangent linear and adjoint of the Weather Research and Forecasting (WRF) Model, Journal of Atmospheric and Oceanic Technology, 30, 1180–1188, 2013b.

Zou, L. and Zhou, T.: Development and evaluation of a regional ocean-atmosphere coupled model with focus on the western North Pacific summer monsoon simulation: Impacts of different atmospheric components, Science China Earth Sciences, 55, 802–815, 2012.