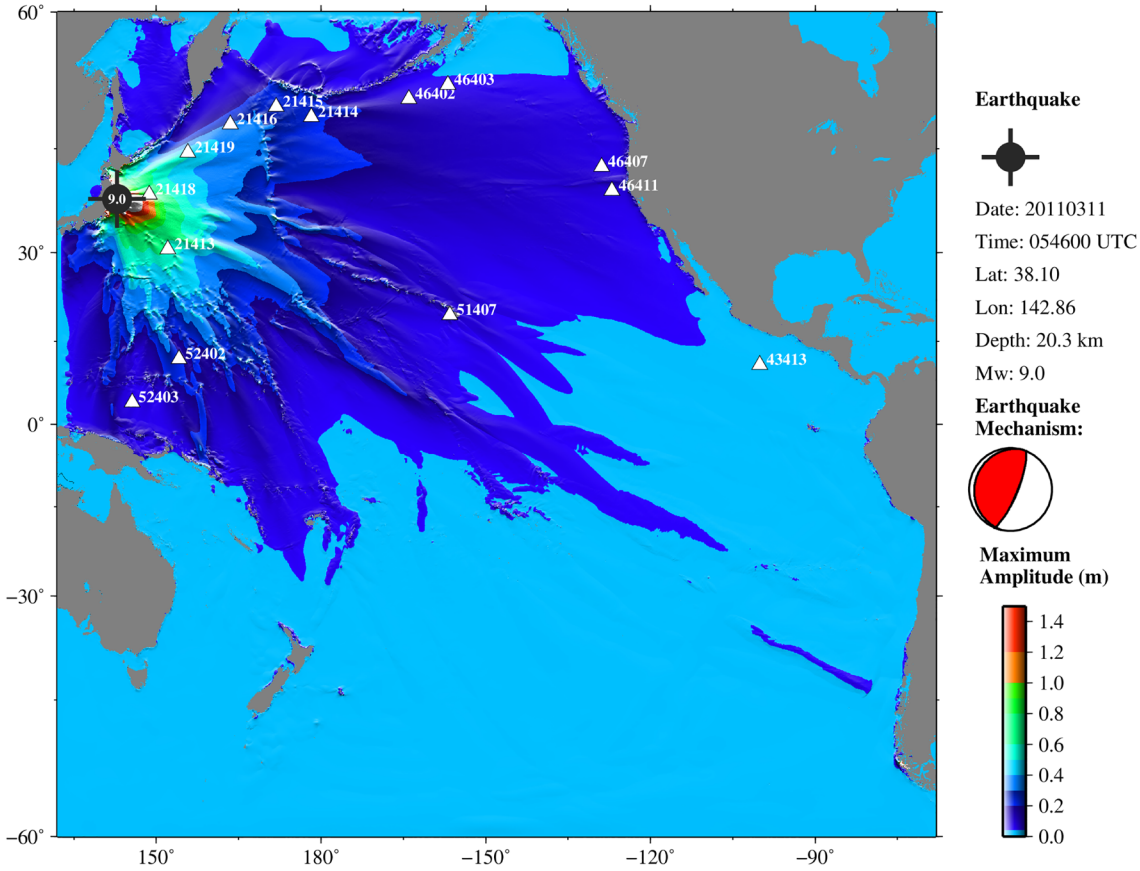

Figure 10. Contour plot of maximum tsunami wave height for Tohoku tsunami on 11 March 2011, modeled by FUNWAVE-GPU. The DART buoys used for comparisons are marked as white triangles.

gradually. Wave breaking-generated vortices are generally confined to the surf zone, and some small-scale vortices are evolved and then shed toward the offshore direction. The vector plots of wave-averaged current reveal offshore-directed rip current is formed and strongest along the rip channels.

Figure 8 illustrates performance of single- and double-GPU code with different mesh grid sizes. Overall, the larger the mesh grid size is, the better the speedup shows. GPU's massive parallelism on computation begins to show its superiority when mesh grid larger than $1,024 \times 1,024$. For modeling tasks with smaller mesh domains of 512×512, due to insufficient occupancy of the CUDA cores, sequential algorithm for tridiagonal systems, and unavoidable overhead, GPU implementation can only achieve a minor speedup of 2.7. This is the case especially for the 2-GPU run, which only shows a speedup of 3.5. The performance exhibits minor improvement with the optimization on concurrent kernels (Figure 8b). The peak performance of GPU is not achieved as suggested by Figure 9. If we measure the GPU performance by a metric of percentage of time that more than one kernel is executing, only 76% and 65% of time that GPU is busy with computational kernels for single- and double-GPU cases, respectively.

For the single-GPU run with concurrent kernels, the metric of speedup ratio climbs up to 7.1 and stay stable above 6.6 as the mesh grid size larger than $1,024,024$. For double-GPU run, a speedup of 14.2 for $4,096 \times 4,096$ suggests a near-linear scalability over the single-GPU run (7.8). However, this perfect scalability cannot be achieved for runs of smaller modeling domain. The reasons are presented in the previous section. When referring to Figure 9, we find the GPU is occupied 97% and 91% of the executing time for the single- and double-GPU runs. For double-GPU run, the global domain is partitioned as two smaller subdomains, the GPU occupancy lowers slightly as expected.

4.3. Case 3: Tohoku Tsunami Modeling Using Spherical Coordinates

In order to validate the CUDA implementation of the code in spherical coordinates, numerical simulation of $M_w = 9.0$ Tohoku tsunami on 11 March 2011 at 05:46 UTC is made with two spatial resolutions