It is essential to document how well the current generation of climate models performs in simulating past climates to have confidence in their ability to project future conditions. We present the first global, in-depth comparison of Pliocene sea surface temperature (SST) estimates from a coupled ocean–atmosphere climate model experiment and a SST reconstruction based on proxy data. This enables the identification of areas in which both the climate model and the proxy dataset require improvement. In general, the fit between model-produced SST anomalies and those formed from the available data is very good. We focus our discussion on three regions where the data–model anomaly exceeds 2 °C. 1) In the high latitude North Pacific, a systematic model error may result in anomalies that are too cold. Also, the deeper Pliocene thermocline may cause disagreement along the California margin; either the upwelling in the model is too strong or the modeled thermocline is not deep enough. 2) In the North Atlantic, the model predicts cooling in the center of a data-based warming trend that steadily increases with latitude from + 1.5 °C to >+ 6 °C. The discrepancy may arise because the modeled North Atlantic Current is too zonal compared to reality, which is reinforced by the lowering of the altitude of the Pliocene Western Cordillera Mountains. In addition, the model's use of modern bathymetry in the higher latitudes may have led the model to underestimate the northward penetration of warmer surface water into the Arctic. 3) Finally, though the data and model show good general agreement across most of the Southern Ocean, a few locations show offsets due to the modern land–sea mask used in the model. Additional considerations could account for many of the modest data–model anomalies, such as differences between calibration climatologies, the oversimplification of the seasonal cycle, and differences between SST proxies (i.e. seasonality and water depth). New SST estimates from data-sparse and regionally important areas will greatly enhance our ability to judge model performance.